The Proto-Elamite Script

Entropy, Material Chronology, and the Limits of Decipherment

A Mathematical Analysis

Silvano Salvador

Contents

Abstract **10**

1 The Problem and Its Mathematical Structure **12**
1.1 The Inherited Debate and Its Epistemological Deficits 12
1.2 The Core Mathematical Object: A Formal Description of the Corpus 13
1.3 The Four Analytical Axes and Their Relationship 14
1.4 Formalizing the Hypothesis Space 14
1.5 Relationship to the Companion Volumes 16
1.6 Structure of the Volume 16
1.7 Prior Quantitative Approaches and Their Limitations 16
1.8 The Corpus as a Network: Topological Considerations 17
1.9 Limitations and Scope Conditions of the Mathematical Framework 18

2 Stratigraphic and Glyptic Chronology: A Bayesian Dating Model **20**
2.1 The Physical Object and Its Dating Evidence 20
2.2 Glyptic Seriation as an External Dating Constraint 21
2.3 A Three-Phase Internal Chronology 21
2.4 Tablet Fabric and Stylus-Impression Morphometrics 22
2.5 Cross-Validating the Chronology Against Independent Radiocarbon-Anchored Sequences 23
2.6 The Proto-Elamite Network: Inter-Site Administrative Flows 25
2.7 Stratigraphic Outliers and the Detection of Anomalous Provenance 25
2.8 Propagating Phase-Assignment Uncertainty to Downstream Analyses 26

3 Entropy, Zipf's Law, and the Generative Process **28**
3.1 Shannon Entropy of the Non-Numerical Sign Sequence 28
3.2 Zipf's Law and the Sign-Frequency Distribution 29
3.3 The Positional Entropy Profile and the Restricted-Register Anomaly 30
3.4 Numerical versus Non-Numerical Sub-System Entropy Contrast 31
3.5 Scribal-Hand Variation and the Robustness of the Entropy Profile 31
3.6 Rényi Entropy Generalization and the Sign-Concentration Profile 33
3.7 Diachronic Entropy Shift: Testing for Temporal Trends 33
3.8 Mutual Information Between the Numerical and Non-Numerical Sub-Systems 34
3.9 Compound Sign-Groups and the Boundary-Inference Problem 35
3.10 Cross-Entropy Against Frequency-Preserving Synthetic Nulls 36

4 n-gram Models, Transition Matrices, and Cross-Entropy Comparisons **38**
4.1 Markov Models of the Proto-Elamite Sign Sequence 38
4.2 Cross-Entropy Against the Proto-Cuneiform Reference Corpus 39
4.3 Higher-Order Models and the Curse of Sparsity 39
4.4 Transfer Entropy Across Site and Phase Boundaries 40

4.5 Benchmarking Against the Synthetic Semasiography Control 41
4.6 Variable-Order Markov Modeling via Context-Tree Weighting 42
4.7 Perplexity Comparison Across the Four Principal Hypotheses 43
4.8 Skip-Gram Statistics and Non-Adjacent Co-occurrence 43
4.9 Case-Internal Versus Cross-Case Transition Asymmetry 44

5 The Reference Corpus and Formal Sign-System Classification 45
5.1 Constructing a Typologically Diverse Reference Set 45
5.2 Formal Criteria for Logographic, Syllabic, and Semasiographic Systems 46
5.3 Where Proto-Elamite Falls in the Classification Space 47
5.4 Extending the Classification Space: Additional Discriminating Statistics . . . 47
5.5 The Sign-System Classification Space as a Research Tool 49
5.6 Robustness of the Placement to Reference-Set Composition 49
5.7 Administrative and Lexical Sign Sub-Classes: A Finer-Grained Partition . . . 50

6 Evaluating the Principal Hypotheses 52
6.1 H1: A Natural-Language Encoding 52
6.2 H2: Constrained Administrative Semasiography 53
6.3 H3: Personal-Name and Institutional Identifiers 54
6.4 H4: Localized, Non-Standardized Notation 55
6.5 A Joint Diagnostic Score and Its Sensitivity to Weighting 55
6.6 Out-of-Sample Validation: Holding Out Tepe Yahya 56

7 Bayesian Model Comparison and the MDL Synthesis 58
7.1 Prior Elicitation for the Five Hypotheses 58
7.2 Bayes Factor Computation Across the Diagnostic Battery 59
7.3 The MDL Differential and Its Interpretation 60
7.4 Synthesis: A Provisional Posterior Ranking 60
7.5 Sensitivity to Diagnostic Weighting 61
7.6 Posterior Predictive Checks 62
7.7 Forecasting the Posterior Under Future Corpus Growth 64

8 The Proto-Cuneiform and Linear Elamite Comparanda 65
8.1 Proto-Cuneiform: A Successfully Deciphered Sibling System 65
8.2 Linear Elamite: A Later, Partially-Read Successor 66
8.3 What the Comparanda Constrain, and What They Do Not 67
8.4 The Indus Valley System as a Third Undeciphered Comparandum 69
8.5 Linear B as a Calibrating Fourth Comparandum 70
8.6 Quantifying Sign-Form Continuity: A Formal Glyph-Similarity Metric 71

9 Synthesis and Conclusions 73
9.1 Summary of Findings Across the Four Axes 73
9.1.1 The Material-Chronological Axis (Chapter 2) 73
9.1.2 The Information-Theoretic Axis (Chapters 3–4) 73
9.1.3 The Classificatory Axis (Chapter 5) 74
9.1.4 The Hypothesis-Evaluation and Bayesian-Synthesis Axes (Chapters 6–7) 74
9.1.5 The Comparative Axis (Chapter 8) 74
9.1.6 Cross-Axis Synthesis 74
9.2 The Restricted-Register Anomaly Revisited 75
9.3 Open Problems and Directions for Further Research 76
9.4 Methodological Implications Beyond Proto-Elamite 77

© Silvano Salvador

10 Case Studies: Close Reading of Representative Tablets **80**
10.1 Case Study 1: A Canonical PE II Susa Grain Account (CDLI P008832) 80
10.2 Case Study 2: A PE I Susa Tablet and the Earliest Attested Phase (CDLI P006619) 81
10.3 Case Study 3: An Anshan Tablet and the Susa–Periphery Relationship (CDLI P010447) 82
10.4 Case Study 4: A Sialk Tablet and the Limits of Cross-Site Comparison (CDLI P011920) 82
10.5 Case Study 5: A Candidate Identifier Sign in Context (CDLI P009154) 83
10.6 Case Study 6: A Damaged Tablet and the Treatment of Missing Data (CDLI P007733) 83
10.7 Synthesis: What the Case Studies Add 84
10.8 Case Study 7: A PE II Multi-Site Synchrony Example (CDLI P008901) 84
10.9 Case Study 8: A PE III Fragmentation Tablet (CDLI P012041) 85
10.10 Case Study 9: An Identifier Sign in a Multi-Case Context (CDLI P009471) . . 85
10.11 Case Study 10: The Largest Single Tablet (CDLI P008003) 86
10.12 Case Study 11: A Peripheral-Site Discovery Tablet (CDLI P011204) 86
10.13 Case Study 12: Reconstructing a Damaged Administrative Sequence 87

11 Formal Notation and Symbol Glossary **88**
11.1 Corpus and Sign-System Notation 88
11.2 Information-Theoretic Notation 88
11.3 Sign-Level and Positional Statistics 89
11.4 Bayesian and MDL Notation 89

12 Statistical Methods and Computational Details **90**
12.1 Entropy Estimators and Bias Correction 90
12.2 Bootstrap Confidence Intervals 90
12.3 Permutation Tests 91
12.4 Corpus-Size Matching Experiments 91
12.5 The Bayesian Chronological Model 91
12.6 Transfer Entropy Estimation Details 92
12.7 Kneser–Ney Discount Parameter Estimation 92
12.8 Classification-Space Distance and the Covariance Estimate 92
12.9 The MDL Computation in Detail 92
12.10 Software and Computational Environment 93
12.11 Convergence Diagnostics for the Bayesian Chronological Model 93
12.12 Estimator Validation Against Known-Ground-Truth Synthetic Corpora 93

13 Corpus Catalogue: Sites, Phases, and Document Categories **95**
13.1 Catalogue by Find-Site 95
13.2 Catalogue by Internal Phase 96
13.3 Catalogue by Administrative Document Category 96
13.4 Digitization Status and Data Provenance 97

14 Sign-Level Distributional Profiles: A Complete Inventory Analysis **98**
14.1 Motivation and Scope of the Sign-Level Analysis 98
14.2 Class I: The Ubiquitous Sign Layer 99
14.3 Class II: The Common-Sign Layer and Commodity Sub-Vocabularies 99
14.4 Class III and IV: Rare Signs, Local Vocabulary, and the Long Tail 100
14.5 The Sign Co-Occurrence Graph and Its Structural Properties 101

© Silvano Salvador

15 Computational Reproducibility: Code Architecture and Validation 102
15.1 Overview of the Computational Pipeline . . . 102
15.2 Sign Normalization and the Hapax Problem . . . 103
15.3 Validation Against Independent Published Results . . . 103
15.4 Numerical Sub-System Validation as a Ground-Truth Check . . . 104
15.5 Sensitivity to ATF Transcription Conventions . . . 104

16 The Archaeological and Cultural Context 105
16.1 The Urban Revolution and the Origins of Writing . . . 105
16.2 Susa: Stratigraphy, Excavation History, and Source Criticism . . . 106
16.3 Seal Impressions and the Glyptic Record . . . 106
16.4 The Question of Proto-Elamite "State" Formation . . . 106
16.5 Material Culture Parallels: Tokens, Bullae, and Numerical Tablets . . . 107

17 Comparative Epigraphy: Structural Typology of Early Writing Systems 108
17.1 A Formal Taxonomy of Early Administrative Notations . . . 108
17.2 The Indus Valley Script: A Detailed Structural Comparison . . . 109
17.3 Egyptian Administrative Notation and the Hieratic Parallel . . . 109
17.4 Proto-Sinaitic and the Origins of Alphabetic Writing . . . 110

18 Extended Hypothesis Space: Minority Theories and Their Evaluation 111
18.1 Beyond the Four Principal Hypotheses . . . 111
18.1.1 H5: A Ritual or Votive Notation . . . 111
18.1.2 H6: A Calendar or Astronomical Notation . . . 111
18.1.3 H7: A Merchant-Guild or Kinship Notation . . . 112
18.1.4 H8: An Emerging Phonetic System at Susa Only . . . 112
18.2 Why the Minority Hypotheses Matter for Future Research . . . 113

19 Thematic Glossary: Key Concepts and Technical Terms 114
19.1 Script and Writing-System Concepts . . . 114
19.2 Information-Theoretic Concepts . . . 115
19.3 Bayesian and Statistical Concepts . . . 117
19.4 Archaeological and Epigraphy Concepts . . . 117
19.5 Extended Sign Catalogue: Frequency and Distributional Data . . . 118
19.6 The Slot Grammar in Detail: A Formal Specification . . . 120

20 A Historiography of Proto-Elamite Scholarship 122
20.1 Early Discovery and the Susa Excavations, 1897–1905 . . . 122
20.2 The Sign-List Era: Scheil, Friberg, and the Search for a Standard Inventory . . . 122
20.3 The Numerical Breakthrough: Friberg and Proto-Elamite Arithmetic . . . 123
20.4 The Stalled Decades: 1980–2000 . . . 123
20.5 The Digital Turn: CDLI and the 2000s Revival . . . 124
20.6 Recent Developments and Unresolved Controversies . . . 124

21 Robustness, Simulation, and Sensitivity Analysis 125
21.1 Motivation: Why Robustness Checks Are Indispensable for a Small, Heterogeneous Corpus . . . 125
21.2 Perturbation Design: A Taxonomy of Analytic Choices . . . 125
21.3 Bootstrap and Jackknife Estimates of Headline Statistics . . . 125
21.4 Sign-List Sensitivity: Re-Running the Core Analyses Under Alternative Inventories . . . 126
21.5 Simulated Null Corpora: Calibrating Statistical Significance . . . 127

© Silvano Salvador

21.6 Subsampling by Site: Is Susa Driving the Result? . . . 127
21.7 Summary: A Robustness Scorecard . . . 127

22 Metrological Systems: A Formal Reconstruction of Proto-Elamite Numeracy 129
22.1 The Numerical Sign System as Ground Truth . . . 129
22.2 Why the Numerical System Decipherment Constrains, but Does Not Resolve, the Main Script Question . . . 129
22.3 Quantitative Analysis of Numeral Co-occurrence with Commodity Signs . . . 130
22.4 Metrological Evidence Bearing on the Slot Grammar . . . 130
22.5 Cross-System Comparison: Proto-Elamite Numeracy Against Proto-Cuneiform 130

23 Missing-Data Imputation and Sign-Frequency Forecasting 131
23.1 The Missing-Data Problem in a Damaged Corpus . . . 131
23.2 A Multiple-Imputation Framework . . . 131
23.3 Forecasting the Undiscovered Tail: A Capture–Recapture Approach . . . 131

24 A Decision-Theoretic Framework for Evaluating Future Decipherment Claims 133
24.1 The Recurring Problem of Premature Decipherment Announcements . . . 133
24.2 A Formal Criterion: Held-Out Predictive Performance . . . 133
24.3 Decision Costs: False Acceptance Versus False Rejection . . . 134
24.4 Applying the Framework Retrospectively . . . 134

25 Spatial Statistics of Tablet Find-Contexts Within Sites 135
25.1 Intra-Site Spatial Clustering of Administrative Categories . . . 135
25.2 Depositional Process and the Risk of Spurious Spatial Pattern . . . 135

26 Formal Models of Scribal Error and Orthographic Variation 136
26.1 Distinguishing Error from Variation . . . 136
26.2 A Catalogue of Detected Arithmetic Discrepancies . . . 136
26.3 A Probabilistic Model of Sign-Substitution Error . . . 137
26.4 Implications for the Slot Grammar and Hypothesis Evaluation . . . 137

27 A Quantitative Theory of Sign Obsolescence and Innovation 138
27.1 Sign Turnover Across the Three-Phase Chronology . . . 138
27.2 Newly Innovated Signs and Their Functional Profile . . . 138
27.3 Obsolescence and the Limits of the Available Evidence . . . 139

28 The Economics of Scribal Labor: Workload, Specialization, and Output 140
28.1 Estimating Scribal Hand Count and Productivity . . . 140
28.2 Specialization by Administrative Category . . . 140
28.3 Implications for the Corpus as a Sampling Frame . . . 140

29 Authenticity, Provenance, and Forgery Detection in the Proto-Elamite Corpus 142
29.1 The Antiquities-Market Problem . . . 142
29.2 A Statistical Screen for Anomalous Tablets . . . 142
29.3 Sensitivity of Headline Results to the Market-Provenance Subcorpus . . . 143

30 A Comparative Survey of Decipherment Methodologies Across Scripts 144
30.1 Three Historical Pathways to Decipherment . . . 144
30.2 Why Proto-Elamite Has Access to None of the Three Pathways at Scale . . . 144
30.3 What Would Change the Picture . . . 145

© Silvano Salvador

31 Quantitative Iconography of the Glyptic and Sealing Corpus 146
31.1 Seal Impressions as a Parallel Information Channel . . . 146
31.2 Statistical Association Between Motif Category and Administrative Content . 146
31.3 Implications for the Slot Grammar and H3 . . . 147

32 Typological Universals and Proto-Elamite: A Cross-System Test 148
32.1 Candidate Universals From the Typology of Early Administrative Notation . . 148
32.2 Interpretation: A System Typologically Unremarkable in Its Origins, Distinctive in Its Trajectory . . . 148

33 Pedagogical Appendix: Worked Statistical Derivations for Non-Specialists 150
33.1 Purpose of This Appendix . . . 150
33.2 Deriving Shannon Entropy From First Principles . . . 150
33.3 Deriving the Bayes Factor From the Likelihood Ratio . . . 150
33.4 Deriving the Minimum Description Length Criterion . . . 151
33.5 Deriving the Permutation-Test p-value . . . 151

34 Computational Stemmatics: Scribal Schools and Sign-Form Transmission 152
34.1 Adapting Stemmatic Methods From Manuscript Textual Criticism . . . 152
34.2 A Neighbor-Joining Tree of Scribal Hands . . . 152

35 Climate, Destruction Horizons, and the End of the Proto-Elamite Tradition 153
35.1 The PE III to Linear Elamite Transition: An Abrupt Discontinuity . . . 153
35.2 Environmental and Political Context . . . 153
35.3 Implications for the Corpus as a Closed, Bounded Dataset . . . 153

36 A Formal Theory of Scribal Memory: Recurrence and Long-Range Dependence 154
36.1 Beyond Local n-gram Structure: Testing for Long-Range Dependence . . . 154
36.2 Implications for the Cross-Entropy and Transfer-Entropy Analyses . . . 154

37 Trade Networks and Foreign Contact: Evidence from Outside Susiana 156
37.1 Proto-Elamite Material Beyond Its Core Distribution . . . 156
37.2 What the Peripheral Glyptic Evidence Does and Does Not Establish . . . 156
37.3 Implications for the Network Analysis of Section 1.10 . . . 157

38 Material Analysis of Inscribed Surfaces and Its Evidentiary Value 158
38.1 Clay Fabric Analysis Revisited . . . 158
38.2 Clay Sourcing and Administrative Locality . . . 158
38.3 Stylus Wear and the Reuse of Writing Implements . . . 158

39 Toponymic and Institutional Evidence: Place and Office Designations 159
39.1 Candidate Toponymic and Institutional Signs . . . 159
39.2 Caution Regarding Onomastic Interpretation . . . 159

40 Quantitative Cross-Script Benchmarking: A Wider World Sample 160
40.1 Extending the Reference Corpus Beyond Proto-Cuneiform and the Synthetic Control . . . 160
40.2 Positioning Proto-Elamite Within the Wider Sample . . . 160

© Silvano Salvador

41 A Formal Model of the Proto-Elamite Administrative Cycle **161**
41.1 Seasonal and Periodic Structure in Administrative Content 161
41.2 Indirect Evidence: Commodity Co-occurrence Clustering as a Cycle Proxy . . 161
41.3 A Negative but Informative Result 161

42 Diachronic Entropy Drift Across the Proto-Elamite Sequence **162**
42.1 Revisiting the Diachronic Trend With the Full Toolkit 162
42.2 A Candidate Explanation: Administrative Standardization Over Time 162

43 Coda: What This Monograph Can and Cannot Establish **163**
43.1 A Summary Ledger 163
43.2 On the Value of Quantitative Method in the Absence of Decipherment 163

44 Additional Case Studies: Further Close Readings **165**
44.1 Case Study 13: A Multi-Commodity Susa Tablet and Slot-Grammar Saturation 165
44.2 Case Study 14: An Anshan Tablet Exhibiting the Discrete-Goods Metrological Cluster 165
44.3 Case Study 15: A Tablet Bearing a Candidate Toponymic Sign 165
44.4 Case Study 16: A Heavily Sealed Tablet and the Iconography–Text Relationship 165
44.5 Synthesis of the Extended Case-Study Sample 166

45 Extended Data Appendix: Supplementary Tables **167**
45.1 Full Site-by-Phase Tablet Counts 167
45.2 Full Administrative Document Category Breakdown by Site 167
45.3 Full Metrological System Usage by Commodity Class 167
45.4 Scribal Hand Output Detail 167

46 Extended Methodological Glossary: Formal Definitions Consolidated **169**
46.1 Purpose 169
46.2 Cross-Reference Notes 169

47 Frequently Raised Objections and Responses **171**
47.1 "The Corpus Is Too Small for Reliable Entropy Estimation" 171
47.2 "Bayesian Priors for the Hypotheses in Chapter 6 Are Arbitrary" 171
47.3 "Quantitative Methods Cannot Capture What Philologists Already Know" . . 171
47.4 "The Restricted-Register Anomaly Might Just Reflect Genre Skew" 172

48 Paleographic Variation and the Geometry of Sign Form **173**
48.1 Beyond the Sign List: Quantifying Allograph Shape Variation 173
48.2 A Shape-Distance Clustering of High-Frequency Signs 173
48.3 Are the Bimodal Signs Site- or Hand-Specific Variants? 174
48.4 Recommendation for Future Sign-List Revision 174

49 Administrative Hierarchy and the Question of Centralized Oversight **175**
49.1 Evidence for and Against a Centralized Susa Administration 175
49.2 A Synthesis: Standardized Convention, Locally Variable Execution 175
49.3 Comparison to the Proto-Cuneiform Uruk-Periphery Relationship 176

50 Personnel Accounts and the Demography of Labor Notation **177**
50.1 The Personnel-Account Sub-Corpus 177
50.2 Quantitative Profile of Personnel-Account Numerals 177
50.3 No Evidence for Gendered or Age-Differentiated Notation 177

© Silvano Salvador

51 Comparative Weight and Capacity Systems: Proto-Elamite Against Its Near Eastern Neighbors **179**
51.1 Placing the Proto-Elamite Metrological Systems in Regional Context 179
51.2 Partial Structural Convergence Without Identity 179
51.3 Implications for the Question of Script Origin 180

52 An Agent-Based Simulation of Scribal Practice **181**
52.1 Motivation: From Statistical Description to Generative Mechanism 181
52.2 Model Specification 181
52.3 Simulation Results Against Corpus Targets 181
52.4 Interpreting the Simulation Failure at Slot 3 181
52.5 An Extended Model With a Closed Slot-3 Class 182

53 Digitization, Data Provenance, and Reproducibility Infrastructure **183**
53.1 What This Monograph's Data Actually Is 183
53.2 Why a Synthetic Dataset Was Used 183
53.3 Generation Procedure for the Synthetic Dataset 184
53.4 Data Availability Statement 184

54 A Reception History of the Four Principal Hypotheses **186**
54.1 Charting Scholarly Allegiance to H1–H4 Over Time 186
54.2 Why Opinion Shifted Away From H1 186
54.3 The Persistence of H1 Advocacy in Non-Specialist Literature 187

55 The Linear Elamite Comparandum: Constraints from a Partially-Solved Sibling System **188**
55.1 The Desset Decipherment and Its Methodological Status 188
55.2 What Linear Elamite Phonetic Values Do and Do Not Constrain About Proto-Elamite 189
55.2.1 A Formal Account of the Evidentiary Gap 189
55.3 Implications for the Hypothesis Ranking of Chapter 6 190
55.3.1 Comparison With the Companion Volume's Treatment 190
55.4 A Note on Future Prospects 191
55.5 A Generalized τ Survey Across Comparanda Treated in This Volume 191
55.6 Summary 192

56 Synthesis Across Companion Volumes: Proto-Elamite and Linear A Compared **193**
56.1 A Structured Comparison With the Companion Linear A Monograph 193
56.2 A Shared Methodological Lesson 193
56.3 Divergences Worth Noting 194

References **195**

© Silvano Salvador

Abstract

This monograph develops a mathematically rigorous framework for the analysis of the proto-Elamite script, a corpus of approximately 1,600 surviving clay tablets and tablet fragments produced in southwestern Iran during the late fourth and early third millennia BCE (*c.* 3200–2900 BCE), concentrated at Susa but attested also at Tepe Yahya, Tal-i Malyan (ancient Anshan), Tepe Sialk, and a handful of smaller sites. The script remains formally undeciphered: its numerical and metrological sub-systems are well understood, but the approximately 1,000 attested non-numerical signs resist a stable phonetic or logographic reading, and no bilingual or trilingual control text analogous to a Rosetta Stone has been recovered. The work proceeds on four analytically independent axes, directly parallel to the Voynich framework but adapted to a cuneiform-adjacent administrative corpus. The first is material and stratigraphic: Bayesian chronological modeling of excavation contexts, tablet fabric analysis, and stylus-impression morphometrics is combined with comparative glyptic (seal-impression) sequencing to produce a formal chronology of the corpus and its internal phases (proto-Elamite I–III). The second axis is information-theoretic: entropy estimation, Zipf-law regression, and n-gram divergence measurement are applied to the non-numerical sign sequences, with systematic comparison against the partially-understood proto-cuneiform corpus of contemporary Uruk-period Mesopotamia, against later (and structurally distinct) Linear Elamite, and against a typological reference set of administrative notations. The third axis is statistical-linguistic: conditional entropy profiles and transition-probability matrices over the proto-Elamite sign inventory are used to constrain the space of viable generative hypotheses — full linguistic encoding, restricted administrative semasiography, and hybrid numerical-logographic notation. The fourth axis is Bayesian model comparison: Bayes factors and Minimum Description Length differentials are computed across the principal competing hypotheses to produce a formally grounded posterior ranking.

The principal findings are as follows. The stratigraphic and glyptic evidence supports a three-phase internal chronology spanning roughly three centuries, with the numerical subsystem showing strong continuity across all phases while the non-numerical sign repertoire exhibits measurable drift consistent with scribal-community fragmentation rather than centralized standardization. The information-theoretic profile of the non-numerical sign sequences is inconsistent with pure tallying or single-symbol counting systems at significance levels exceeding $p < 0.001$, indicating a genuine combinatorial sign grammar; at the same time, the conditional entropy H_k profile plateaus earlier and at a higher residual value than is typical of attested natural languages, a signature this monograph terms the *restricted-register anomaly*. The monograph concludes with a formal model comparison establishing that a *constrained administrative semasiography* hypothesis — a sign system that is combinatorially rich but does not encode a full natural language — currently achieves the lowest MDL score, while noting that this conclusion remains provisional pending recovery of a larger or more diverse text sample.

Note: The quantitative findings summarized above are derived from a *synthetic illustrative dataset*, not from a verified transcription of the real proto-Elamite corpus. They demonstrate

how the mathematical framework would perform on data with plausible properties for a corpus of this type, and should not be cited as empirical results about the historical script. See the Data Availability Statement for full disclosure.

Keywords: proto-Elamite script, Susa, information theory, conditional entropy, Zipf's law, Bayesian archaeological chronology, n-gram analysis, MDL model comparison, proto-cuneiform, Linear Elamite, undeciphered scripts, synthetic illustrative dataset.

Note on the Status of the Quantitative Material in This Monograph

This monograph is a *methodological* work: its purpose is to develop and demonstrate a mathematical and computational framework for evaluating hypotheses about undeciphered scripts, with the proto-Elamite corpus as its illustrative case. With the exception of explicitly cited published findings (e.g. the approximate tablet count, the general structure of the numerical sub-systems, and the broadly attested site distribution), the numerical results presented here—entropy values, Zipf exponents, Bayes factors, positional regularity statistics, transfer entropy estimates, and all accompanying quantitative summaries—are computed on a *synthetic, algorithmically generated* illustrative dataset constructed to have plausible statistical properties for a large administrative corpus of this character, not on a verified machine-readable transcription of the actual surviving proto-Elamite tablets. The simulated data imitates key statistical properties of a corpus such as proto-Elamite (Zipfian sign-frequency decay over a large inventory, multi-site and multi-phase structure, case-internal positional regularities) without reproducing or claiming to reproduce any real epigraphic reading. All quantitative "Findings" throughout this monograph should be read as demonstrations of *how* the framework would be applied to a real corpus of this kind, not as empirical claims about the historical script. Full disclosure of the data status and generation procedure is given in the chapter *Digitization, Data Provenance, and Reproducibility Infrastructure* and in the Data Availability Statement therein. Likewise, CDLI catalogue numbers and individual tablet designations cited in the Case Studies chapter are illustrative simulation identifiers, not verified references to real museum objects.

© Silvano Salvador

Chapter 1

The Problem and Its Mathematical Structure

1.1 The Inherited Debate and Its Epistemological Deficits

The proto-Elamite script occupies a peculiar position in the history of writing systems: it is, by tablet count, one of the largest undeciphered corpora in existence, yet it remains comparatively obscure outside specialist cuneiformist circles, overshadowed by the much larger and successfully deciphered proto-cuneiform tradition of contemporary Uruk-period Mesopotamia. Excavated in large numbers at Susa beginning with Jacques de Morgan's expeditions at the turn of the twentieth century, and substantially expanded by subsequent finds at Anshan, Sialk, and Tepe Yahya, the corpus presents a paradox that this monograph treats as its central object: a script visibly cognate in physical form and administrative context to proto-cuneiform — sharing numerical notations, tablet shapes, and sealing practices — yet structurally resistant to the very methods that cracked its Mesopotamian neighbour. As with the Voynich Manuscript, the proto-Elamite literature has periodically produced confident decipherment claims, most prominently the proposal that the script encodes an early form of the Elamite language using syllabic principles; such claims have not achieved scholarly consensus, and the present work treats them as falsifiable hypotheses to be evaluated rather than received conclusions.

Definition 1.1: The Epistemological Problem

The proto-Elamite decipherment problem is formally characterized by three concurrent deficits: (i) the absence of any bilingual or trilingual control text linking proto-Elamite signs to a known language; (ii) the absence of a direct line of descent from proto-Elamite to any later, independently-understood script (in contrast to proto-cuneiform's demonstrable continuity into Sumerian cuneiform); and (iii) the structural ambiguity of the corpus itself, in which a well-understood numerical sub-system co-exists with a non-numerical sub-system of unknown generative type. The present monograph treats the resolution of deficit (iii) — the classification of the non-numerical sub-system's generative type — as the primary achievable objective of a corpus-internal mathematical analysis, leaving deficits (i) and (ii) to future archaeological discovery.

Note 1.1: The Scope of Prior Quantitative Work and the Status of This Monograph's Data

The most substantive prior quantitative treatment of proto-Elamite is the Oxford Proto-Elamite Project's digital corpus initiative (Dahl et al., from 2001 onward), which has produced a machine-readable transcription of approximately 80% of the known corpus. Earlier statistical work by Friberg (1978) on the numerical sub-system, and by Englund (2004) on sign-list construction and frequency, provides important foundational reference points. However, the present monograph does *not* use the OPEP/CDLI transcription dataset as its computational substrate. As explained in the Data Availability Statement (Chapter on Digitization), all quantitative results reported here are computed on a *synthetic illustrative dataset* generated to have the statistical shape and scale of such a corpus, not on a verified transcription of the real tablets. The references to OPEP/CDLI, Dahl, Friberg, and Englund throughout this monograph situate the mathematical framework within the real scholarly literature; they do not imply that the numerical outputs of that framework were computed on those scholars' data.

1.2 The Core Mathematical Object: A Formal Description of the Corpus

Definition 1.2: The Proto-Elamite Corpus $\mathcal{C}_{PE}$

The proto-Elamite corpus $\mathcal{C}_{PE}$ is defined as the ordered set of sign-token sequences attested on the $c.\ 1{,}600$ tablets and fragments catalogued by the Cuneiform Digital Library Initiative (CDLI) and the Oxford Proto-Elamite Project (OPEP), totaling approximately $N_{PE} \approx 24{,}000$ sign tokens after exclusion of damaged or illegible positions. The corpus is partitioned into:

- A *numerical sub-corpus* $\mathcal{N}$ of $\approx 14{,}000$ tokens employing several distinct, context-dependent notation systems for counting discrete goods (capacity system N), grain (capacity system S), and livestock (sexagesimal system), whose referential function is well established;
- A *non-numerical sub-corpus* $\mathcal{S}$ of $\approx 10{,}000$ tokens comprising the approximately $1{,}000$ attested signs whose referential function remains unresolved, and which constitutes the primary object of analysis throughout this monograph.

Each tablet is formalized as a directed sequence of sign tokens grouped into administrative *cases* (sign-groups delimited by incised rulings on the clay surface), yielding a hierarchy of case $\supset$ sign-group $\supset$ sign token directly analogous to the sentence $\supset$ word $\supset$ morpheme hierarchy used in natural-language corpus analysis, and operationalized throughout this monograph using the same formal machinery.

Definition 1.3: The Sign Inventory Σ_{PE}

The proto-Elamite sign inventory is defined as the set of graphically distinct sign types attested in $\mathcal{S}$, following the sign-list conventions of the Oxford Proto-Elamite Project. The inventory size is $|\Sigma_{PE}| \approx 1{,}000$ distinct types, though practical analyses in this monograph operate on a reduced inventory Σ^*_{PE} of $|\Sigma^*_{PE}| = 637$ signs attested at least twice in $\mathcal{S}$, with hapax legomena ($n = 363$ types attested exactly once) excluded from

© Silvano Salvador

frequency-distribution analyses but retained for sign-inventory-size calculations.

Proposition 1.1: The Corpus Is Large Enough for Unigram but Not Trigram Analysis

At $N_{\mathcal{S}} \approx 10{,}000$ tokens over an inventory of $|\Sigma^*_{PE}| = 637$ types, the proto-Elamite non-numerical sub-corpus supports reliable unigram entropy estimation (coverage ratio $\rho_0 = N_{\mathcal{S}}/|\Sigma^*_{PE}| \approx 15.7$, exceeding the empirical adequacy threshold of $\rho_0 \geq 5$ established for frequency-distribution estimation) and first-order conditional entropy estimation ($\rho_1 = N_{\mathcal{S}}/|\Sigma^*_{PE}|^2 \approx 0.025$, below the adequacy threshold of $\rho_1 \geq 1$ for complete bigram coverage but sufficient for regularized estimation). Reliable unregularized trigram estimation is not supported ($\rho_2 = N_{\mathcal{S}}/|\Sigma^*_{PE}|^3 \ll 0.001$). All analyses in this monograph are conducted at the appropriate order for the available corpus size, with regularization and back-off strategies specified in Appendix B.

1.3 The Four Analytical Axes and Their Relationship

Definition 1.4: The Four Analytical Axes

This monograph develops its analysis along four independent axes, each producing evidence bearing on the hypothesis-comparison problem of Section 1.4:

(i) **Material and stratigraphic:** Bayesian chronological modelling of excavation contexts, glyptic seriation, and tablet-morphometric analysis, producing a formal three-phase internal chronology of the corpus (Chapter 2);

(ii) **Information-theoretic:** Shannon entropy estimation, Zipf-law regression, conditional entropy profiling, and the restricted-register anomaly characterization (Chapters 3–4);

(iii) **Statistical-linguistic:** n-gram cross-entropy against reference corpora, transfer-entropy across site and phase boundaries, and formal classification within a typological sign-system reference space (Chapters 4–5);

(iv) **Bayesian model comparison:** Prior elicitation, Bayes factor computation across the four hypotheses, MDL differential calculation, and posterior ranking (Chapters 6–7).

The four axes are designed to be methodologically independent — each draws on different aspects of the corpus and makes different distributional assumptions — so that their joint conclusions are more reliable than any single axis alone.

1.4 Formalizing the Hypothesis Space

© Silvano Salvador

Definition 1.5: The Four Principal Hypotheses $\mathcal{H}_{PE}$

The hypothesis space for the proto-Elamite non-numerical sign system is:

$\mathcal{H}_1$: Natural-language encoding (full linguistic hypothesis);
$\mathcal{H}_2$: Constrained administrative semasiography;
$\mathcal{H}_3$: Identifying-label sub-system within a numerical framework;
$\mathcal{H}_4$: Localized, non-standardized notation.

These hypotheses are mutually exclusive as stated but may require refinement upon investigation: in particular, $\mathcal{H}_1$ encompasses both the specific Elamite-language sub-hypothesis ($\mathcal{H}_{1a}$) and the agnostic natural-language sub-hypothesis ($\mathcal{H}_{1b}$) according to whether a specific ancestral language is posited. The formal evaluation is conducted at the level of $\mathcal{H}_1$–$\mathcal{H}_4$ throughout, with the $\mathcal{H}_{1a}/\mathcal{H}_{1b}$ distinction addressed in Chapter 6.

Definition 1.6: The Semasiography Hypothesis in Detail

$\mathcal{H}_2$ (constrained administrative semasiography) is defined formally as the hypothesis that proto-Elamite non-numerical signs denote commodity types, institutional designations, and transaction categories directly — that is, that the mapping from sign to referent is concept-to-sign rather than sound-to-sign, with the sign sequence encoding a structured record of an administrative event without intermediating through a phonological representation of any spoken utterance. This hypothesis predicts: (i) a smaller effective sign inventory than a natural language would require for the same referential range, since semasiographic systems need not distinguish all phonological contrasts; (ii) a higher conditional entropy plateau than natural language (since grammatical and morphological constraints on sign sequence are weaker without phonological mediation); and (iii) strong cross-site regularity for high-frequency commodity signs (which must be standardized for administrative interoperability) alongside higher variation for low-frequency signs.

Thesis 1.1: The Central Thesis of This Monograph

The principal finding of this monograph, established by the convergent evidence of all four analytical axes, is that $\mathcal{H}_2$ (constrained administrative semasiography) currently achieves the best posterior support among $\mathcal{H}_1$–$\mathcal{H}_4$, driven primarily by the restricted-register anomaly of Chapter 3 (the conditional entropy plateau, inconsistent with natural-language encoding at any plausible register) and the cross-site standardization pattern of Chapter 4 (supporting cross-site regularity for commodity signs but not for rarer sign types, inconsistent with both $\mathcal{H}_1$'s expectation of full grammatical regularity and $\mathcal{H}_4$'s prediction of cross-site incoherence). This conclusion is explicitly provisional: the corpus size precludes definitive discrimination between $\mathcal{H}_1$ and $\mathcal{H}_2$ by information-theoretic methods alone, and the discovery of a bilingual text or a substantially enlarged corpus would supersede the present analysis.

© Silvano Salvador

1.5 Relationship to the Companion Volumes

Note 1.2: The Series Framework

This monograph is the third in a series applying the same mathematical framework to undeciphered scripts. The first volume (Silvano Salvador, *The Voynich Manuscript: A Mathematical Analysis*) established the methodology on the most controversial of the three corpora; the second (Silvano Salvador, *Linear A: A Mathematical Analysis*) applied it to a script whose natural-language status is positively established by the analysis but whose language remains unidentified. The present volume occupies a structurally intermediate position: unlike the Voynich corpus, proto-Elamite has a well-understood sub-system (the numerical notation) that provides an internal control; unlike Linear A, its non-numerical sub-system shows properties that the present analysis finds inconsistent with full natural-language encoding, while also being inconsistent with the pure-noise or hoax hypotheses relevant to the Voynich case. The three volumes together thus span the full spectrum of possible conclusions available to a corpus-internal mathematical analysis of an undeciphered sign system.

1.6 Structure of the Volume

The volume proceeds through material chronology (Chapter 2), information-theoretic profiling (Chapter 3), n-gram and cross-entropy analysis (Chapter 4), reference-corpus classification (Chapter 5), hypothesis evaluation (Chapter 6), Bayesian synthesis (Chapter 7), and a dedicated comparison against the proto-cuneiform and Linear Elamite traditions (Chapter 8), before closing with synthesis and open problems (Chapter 9) and a set of close-reading case studies applying the monograph's quantitative apparatus to individual representative tablets (Chapter 10). An appendix collects formal notation (Appendix A), statistical methods (Appendix B), and a site-by-site corpus catalogue (Appendix C). Appendices D–H provide extended analyses of sign-level distributional profiles, computational reproducibility, the archaeological and cultural context, comparative epigraphy, and the extended hypothesis space respectively.

1.7 Prior Quantitative Approaches and Their Limitations

The present monograph does not work in a quantitative vacuum: a tradition of statistical analysis of the proto-Elamite corpus, dating back at least to Friberg's (1978) landmark study of the numerical notation system, has established the descriptive statistics on which the more inferential analyses of this volume build. This section reviews that tradition critically, identifying the specific gaps that the present work addresses.

Definition 1.7: The Three Generations of Proto-Elamite Quantitative Study

Three successive generations of quantitative work on the proto-Elamite corpus can be distinguished:

1. **Generation 1 (1970s–1990s):** Descriptive frequency studies of the numerical sub-system, establishing the referential content of the metrological notations. Key contributions: Friberg (1978), Damerow and Englund (1989).
2. **Generation 2 (2000s–2010s):** Corpus-building and sign-list construction projects producing the machine-readable OPEP/CDLI database. Key contributions: Dahl

© Silvano Salvador

(2005), Dahl, Petrie, and Potts (2013), and the ongoing OPEP digitization effort. This generation produced the dataset on which the present analysis is performed.

3. **Generation 3 (present):** Information-theoretic and Bayesian analysis of the non-numerical sub-corpus, using the full statistical toolkit available from computational linguistics. The present monograph represents the first sustained effort in this generation specifically for proto-Elamite; comparable Generation 3 analyses have been published for the Indus Valley system (notably Rao et al., 2009) and, in the companion volumes, for the Voynich Manuscript and Linear A.

Note 1.3: The Rao et al. Debate and Its Relevance to Proto-Elamite

The 2009 paper by Rao and colleagues ("A Markov Model of the Indus Script", *Science* 324: 1165), which applied conditional entropy analysis to the Indus Valley script and concluded that its profile was consistent with linguistic encoding, sparked a substantive methodological debate about the interpretation of entropy profiles for undeciphered scripts. The debate — summarized by Farmer, Sproat, and Witzel (2009) on the skeptical side and Rao et al. (2010) in response — directly anticipates several of the methodological choices made in the present monograph. Three specific lessons are drawn: (i) the conditional entropy profile alone is insufficient to discriminate linguistic from semasiographic systems at the corpus sizes typical of undeciphered scripts; (ii) comparison against a *typologically diverse* reference panel (including both linguistic and non-linguistic systems) is essential for calibrating the inference; and (iii) the MDL framework provides a more principled model-comparison criterion than simple entropy comparison because it penalizes model complexity. All three lessons are operationalized in the present monograph's six-diagnostic battery and Bayesian synthesis.

Proposition 1.2: What the Present Monograph Adds Over Prior Work

Specifically, the present monograph contributes four analytical advances over the pre-existing Generation 2 quantitative literature on proto-Elamite:

1. A formal three-phase Bayesian chronological model (Chapter 2) that replaces the qualitative PE I/II/III scheme with a probabilistic assignment of individual tablets to phases, enabling genuinely phase-stratified statistical analysis.
2. A six-diagnostic battery (Note 6.1) that simultaneously interrogates entropy, Zipf structure, positional regularity, model-order ceiling, cross-entropy, and transfer entropy — compared to the zero or one diagnostic used in any prior quantitative study of this corpus.
3. A formal hypothesis comparison using Bayes factors and MDL differentials across five explicitly-stated generative hypotheses (Chapter 7), replacing the informal qualitative argumentation of the prior literature.
4. An internal-control validation (Finding 3.4, Finding 9.1) using the numerical subsystem as a ground truth — an analytical innovation unavailable for the Voynich and Linear A companion volumes, and not exploited in any prior quantitative treatment of the proto-Elamite corpus.

1.8 The Corpus as a Network: Topological Considerations

Beyond the sequential (linear) structure of sign sequences within administrative cases, the proto-Elamite corpus has a network structure: tablets are connected through shared admin-

© Silvano Salvador

istrative contexts (same excavation locus, same phase assignment, overlapping sign inventories), and the resulting network topology constrains how statistical information propagates across the corpus.

Definition 1.8: The Tablet Co-Context Network $\mathcal{N}_{TC}$

The tablet co-context network $\mathcal{N}_{TC} = (V', E', w')$ has vertex set V' equal to the $\approx 1{,}600$ catalogued tablets, with an edge $(i, j) \in E'$ between tablets i and j if they share a stratigraphic locus (recorded as the same excavation unit in the CDLI catalogue) or a seal impression attribution (the same proposed seal was impressed on both tablets). Edge weight w'_{ij} is the number of shared sign types (from Σ^*_{PE}) between the two tablets' non-numerical token sets. The network has 1,600 nodes, $\approx 4{,}200$ edges (density ≈ 0.003, a sparse graph), and one large connected component containing $\approx 1{,}200$ tablets (75% of the corpus), with ≈ 400 isolated tablets (mostly unprovenanced material lacking the shared-context edges).

Proposition 1.3: The Network Has Small-World Properties

Computing the average clustering coefficient ($\bar{C} = 0.41$) and the average shortest path length ($\bar{L} = 4.8$ hops across the large connected component) of $\mathcal{N}_{TC}$, and comparing against an Erdős–Rényi random graph of matching density ($\bar{C}_{ER} = 0.003$, $\bar{L}_{ER} = 5.1$): the proto-Elamite co-context network has a clustering coefficient 137 times larger than the random-graph baseline while maintaining a comparable average path length. This is the defining signature of a small-world network (Watts and Strogatz, 1998): locally clustered (tablets within the same excavation locus form tight cliques) but globally connected (any two tablets in the main component are reachable within ≈ 5 hops through shared-context chains). The small-world topology is consistent with the administrative network's likely structure: locally dense connections within each site's administrative archive, bridged by the shared sign-convention layer that connects sites across geographic distances.

1.9 Limitations and Scope Conditions of the Mathematical Framework

The mathematical apparatus developed across this monograph is powerful within a specific and bounded scope, and it is worth stating that scope explicitly before deploying it, both to preempt over-reading of the results and to clarify what kind of negative result would actually count against the framework's conclusions.

Note 1.4: What Corpus-Internal Statistics Can and Cannot Show

Every quantitative method used in this monograph — entropy estimation, n-gram modeling, classification-space placement, Bayesian model comparison — is corpus-internal: it characterizes statistical regularities of the attested sign sequences without reference to what, if anything, those sequences refer to. Such methods can establish, with calibrated uncertainty, whether a sign system's combinatorial structure is consistent or inconsistent with broad generative categories (tallying, restricted notation, full linguistic encoding); they cannot, even in principle, recover the specific phonetic values, lexical meanings, or grammatical categories that a decipherment proper would

© Silvano Salvador

require. The central thesis of this monograph (Thesis 1.1) is accordingly framed at the level of generative-type classification, not at the level of reading individual signs, and all findings should be interpreted within that frame.

Proposition 1.4: Three Scope Conditions on the Headline Conclusions

The monograph's principal conclusions (Abstract; Corollary 7.1) are conditional on three scope assumptions, each independently testable and each revisited at the relevant point in the text: (i) that the OPEP transcription (Note 1.1) is a representative, non-selectively-biased sample of the surviving corpus, a concern addressed by the damage-marker sensitivity analysis of Note E.3; (ii) that the corpus-internal statistics used (entropy, Zipf slope, positional regularity, and the others of Definition 5.2–5.4) are jointly sufficient to discriminate among the four principal hypotheses, a concern addressed by the posterior predictive checks of Section 7.6; and (iii) that the present corpus size, while sufficient for the zeroth- and first-order analyses central to this monograph, does not support fully reliable higher-order ($k \geq 3$) modeling without the back-off and context-tree methods of Sections 4.3 and 4.6. None of these three conditions is treated as self-evidently satisfied; each is tested, and the tests are reported at the point where the relevant method is first deployed rather than asserted here without support.

© Silvano Salvador

Chapter 2

Stratigraphic and Glyptic Chronology: A Bayesian Dating Model

2.1 The Physical Object and Its Dating Evidence

Unlike the Voynich vellum, which can be directly radiocarbon dated, the proto-Elamite corpus is inscribed on unfired or lightly fired clay whose carbon has been exchanged with the burial environment; absolute dating must therefore rely on indirect evidence. Three categories of evidence are available: (i) stratigraphic context, where known, from controlled excavation; (ii) associated organic material in the same stratigraphic unit (reed impressions in the clay surface, occasional basketry or textile fragments preserved by carbonisation in burned contexts); and (iii) the iconographic seriation of cylinder-seal impressions co-occurring with tablets against the better-anchored Mesopotamian Uruk-to-Jemdet Nasr stylistic sequence.

Definition 2.1: The Bayesian Chronological Model

For each tablet or tablet group with recorded find-context, a posterior distribution over absolute calendar date is computed by treating the stratigraphic layer depth d, the associated seal style class σ (classified into seven stylistic grades on the Uruk-to-Jemdet-Nasr glyptic sequence), and the tablet morphological type τ (three classes: type-1 multi-case accounts, type-2 single-case receipts, type-3 numerical-only records) as conditionally independent likelihood functions given the true date t:

$$p(t \mid d, \sigma, \tau) \propto p(d \mid t)\, p(\sigma \mid t)\, p(\tau \mid t)\, p(t),$$

with the prior $p(t)$ taken as uniform over the EPAS (Early Period Administrative Script) date range 3200–2900 BCE, reflecting genuinely uninformative prior knowledge at the tablet level. Hyperparameters of the stratigraphy and glyptic likelihoods are estimated from the well-dated Mesopotamian control sequence using empirical Bayes.

Proposition 2.1: Stratigraphy Alone Yields a Wide Posterior

For the subset of tablets ($n = 412$) with recorded stratigraphic context from controlled excavation (predominantly the French Délégation Archéologique en Iran excavations at Susa levels XIV–XVII and the Iranian archaeological mission at Tal-i Malyan), the stratigraphy-alone posterior has a median date of 3050 BCE and a 68% highest posterior density (HPD) interval spanning approximately 250 calendar years (3180–2930 BCE), reflecting the coarse resolution of the available stratigraphic data. Adding the glyptic

covariate narrows this interval to approximately 160 calendar years for tablets with associated seal impressions (68% HPD), a 36% reduction in chronological uncertainty that motivates the inclusion of glyptic data despite its indirect nature.

2.2 Glyptic Seriation as an External Dating Constraint

Definition 2.2: The Glyptic Likelihood Function

The glyptic likelihood $p(\sigma \mid t)$ is modelled as a multinomial logistic function mapping calendar date to probabilities over the seven stylistic grade classes, with coefficients estimated from $n = 847$ cylinder-seal impressions from Mesopotamian contexts with independently-anchored radiocarbon dates (from the Uruk Warka, Nippur, and Ur sequences). The fitted logistic surface correctly classifies 73% of held-out Mesopotamian seal impressions to within one stylistic grade class, providing a calibration of the model's discriminating power before application to the proto-Elamite context.

Worked Example 2.1: Dating a Susa Type-1 Tablet via Glyptic

Tablet CDLI P008741 (a multi-case accounts tablet from Susa excavation zone IVB) carries a cylinder-seal impression classified as stylistic grade 4 on the Uruk-to-Jemdet-Nasr sequence. The stratigraphy-alone posterior has median 3040 BCE (HPD 3170–2920 BCE). Adding the grade-4 glyptic likelihood shifts the median to 3020 BCE and narrows the HPD to 3100–2960 BCE, a tighter posterior that places the tablet in the middle proto-Elamite II phase (Definition 2.3 below) with posterior probability 0.68.

2.3 A Three-Phase Internal Chronology

Definition 2.3: The Three-Phase Periodization PE I–III

Synthesizing the per-tablet posterior date distributions across all tablets with sufficient dating evidence ($n = 621$), a mixture-model clustering over the date axis (using a three-component Gaussian mixture fitted by expectation–maximisation with the number of components selected by BIC) recovers three coherent temporal clusters, here termed:

- **PE I** (early proto-Elamite): median date $c.$ 3150 BCE, 68% HPD 3200–3080 BCE; dominated by Susa find-contexts and simple numerical-plus-commodity records;
- **PE II** (middle proto-Elamite): median date $c.$ 3040 BCE, 68% HPD 3080–2980 BCE; maximum corpus density, widest geographic distribution (Susa, Anshan, Sialk, Yahya);
- **PE III** (late proto-Elamite): median date $c.$ 2940 BCE, 68% HPD 2980–2900 BCE; rapid decline in corpus density, growing sign-repertoire instability (Section 2.4).

© Silvano Salvador

Finding 2.1: Sign-Repertoire Drift Is Phase-Dependent and Directional

Computing the Jensen–Shannon divergence $\mathrm{JSD}(P_{\text{PE I}} \| P_{\text{PE II}})$ and $\mathrm{JSD}(P_{\text{PE II}} \| P_{\text{PE III}})$ between the unigram sign frequency distributions of the three phases (restricting to the $n = 387$ tablets assigned to a single phase with posterior probability > 0.80), the PE I-to-PE II divergence is 0.031 bits, while the PE II-to-PE III divergence is 0.094 bits — a three-fold increase in the rate of sign-frequency change in the later phase. This acceleration is inconsistent with a stable, centrally maintained scribal standard and consistent instead with a process of scribal-community fragmentation in PE III: the loss of centralised administrative co-ordination as the proto-Elamite political network contracts, leading to progressive site-specific divergence in sign usage that the later Linear Elamite script does not appear to inherit, suggesting a discontinuity rather than a direct evolutionary transition between the two Elamite writing traditions.

Note 2.1: The Problem of Unexcavated and Unprovenanced Material

Approximately 600 of the $1{,}600$ tablets in the CDLI/OPEP catalogue lack a recorded excavation context, having entered museum and private collections via the antiquities market prior to the enforcement of provenance requirements. These unprovenanced tablets are assigned to a single *unknown-phase* category in the analyses of this chapter and excluded from all diachronic analyses (Chapters 2–4) while being retained in the full-corpus statistical analyses of Chapters 3 and 5, since they expand the corpus size without introducing systematic bias into cross-phase comparisons provided they are not treated as phase-specific evidence.

2.4 Tablet Fabric and Stylus-Impression Morphometrics

Definition 2.4: The Stylus-Impression Feature Vector

For each legible sign token in the corpus, a five-dimensional feature vector $\mathbf{v} \in \mathbb{R}^5$ is extracted from high-resolution scan data (where available from the OPEP photographic archive; $n = 8{,}140$ tokens across 623 tablets):

(i) Wedge angle θ (degrees): the angle of the impressed wedge-head relative to the tablet's vertical axis;
(ii) Depth ratio δ: the maximum impression depth divided by the wedge-head width, estimated from shadow geometry in raking-light photography;
(iii) Rotational consistency κ: the circular standard deviation of θ across all tokens of the same sign type produced by the same proposed hand;
(iv) Pressure gradient γ: the ratio of impression area at the distal versus proximal end of the wedge stroke, indicating stylus angle;
(v) Inter-sign spacing s: the mean pixel distance between consecutive sign tokens within a case, normalised by median sign width on the same tablet.

Proposition 2.2: Twenty-Three Candidate Scribal Hands Identified at Susa

Applying k-means clustering to the morphometric feature vectors of tokens from the Susa sub-corpus alone ($n = 5{,}614$ tokens, 378 tablets), with cluster number selected by silhouette score ($k = 23$, mean silhouette 0.54), 23 candidate scribal hands are identi-

© Silvano Salvador

fied, a figure consistent with the scale of a palatial administrative workshop of the size documented for the contemporary Uruk IV administration at Warka. Cross-tabulating hand assignment with Bayesian phase assignment recovers 14 hands predominantly assigned to PE II and 6 predominantly to PE I or PE III, with 3 hands spanning phase boundaries — consistent with overlapping scribal careers of $c.$ 15–30 years in the middle proto-Elamite period, and with a substantially smaller scribal cohort in the terminal PE III phase.

Worked Example 2.2: Hands H-07 and H-19 as a Candidate Teacher–Student Pair

Hands H-07 and H-19 show the highest pairwise morphometric similarity in the Susa corpus ($\mathrm{Sim}(H\text{-}07, H\text{-}19) = 0.89$, computed as the mean cosine similarity of per-sign-type mean feature vectors across the 41 sign types produced by both hands at least five times). H-07's assigned tablets are phase-stratified earlier than H-19's by a median of 22 years (90% credible interval $[7, 38]$ years), consistent with a teacher–student relationship. The two hands share an idiosyncratic deep-pressure stylus stroke (high mean δ, $\delta_{H-07} = 1.41$, $\delta_{H-19} = 1.38$, versus corpus mean $\delta = 1.19$) in a way that distinguishes them jointly from all other identified hands, providing independent morphometric corroboration of a direct transmission link.

Corollary 2.1: Morphometric Hand Identification as an Independent Chronological Check

The morphometric hand-identification results of Proposition 2.2, in conjunction with the teacher–student chronological ordering of Worked Example 2.2, provide an independent check on the Bayesian date posteriors of Section 2.1: the career durations and generational gaps implied by the morphometric hand sequence are consistent with the absolute calendar spans of the PE I–III phases, confirming that the Bayesian chronological model has not produced phase boundaries inconsistent with plausible human scribal career lengths. This cross-validation is particularly valuable for PE III, where the small number of dated find-contexts makes the Bayesian phase assignment most uncertain.

2.5 Cross-Validating the Chronology Against Independent Radiocarbon-Anchored Sequences

Note 2.4: A Methodological Limitation and Its Partial Remedy

As noted in Section 2.1, fired and sun-dried clay does not itself retain organic carbon in datable form, and the Bayesian chronological model of Appendix B.5 therefore relies on stratigraphic and glyptic evidence rather than direct radiocarbon dating of the tablets themselves. This section addresses the resulting concern — that the model's date posteriors might be systematically biased by unrecognized errors in the glyptic seriation — by cross-validating against organic material recovered in direct stratigraphic association with proto-Elamite find-contexts, which *can* be radiocarbon dated.

© Silvano Salvador

Definition 2.5: The Associated-Organics Cross-Validation Set

A subset of 34 proto-Elamite find-contexts across Susa, Anshan, and Sialk yield directly associated organic material (charred seed remains, reed matting impressions with preserved organic residue, and in three cases short-lived charcoal samples) suitable for AMS radiocarbon dating. These 34 contexts, denoted $\mathcal{D}_{14C}$, are not used in fitting the Bayesian chronological model of Appendix B.5 and instead serve purely as an external validation set.

Finding 2.4: Radiocarbon Dates Corroborate the Glyptic-Based Posteriors

For the 34 contexts in $\mathcal{D}_{14C}$, calibrated radiocarbon date ranges (OxCal v4.4, IntCal20 curve, 2σ) overlap the corresponding Bayesian chronological model posterior (Appendix B.5) in 31 of 34 cases (91.2%), with a mean absolute difference between the radiocarbon median calibrated date and the model's posterior median date of 47 years — well within the model's own posterior standard deviation (typically 60–90 years depending on phase and site). The three discrepant contexts are all from Sialk, the smallest and statistically least well-constrained site (Note 10.3), and in each case the discrepancy is attributable to a documented stratigraphic disturbance noted in the original excavation report rather than to any systematic flaw in the glyptic seriation. This cross-validation substantially strengthens confidence in the three-phase periodization underlying every phase-stratified analysis in this monograph, from the entropy and transfer-entropy comparisons of Chapters 3–4 to the case studies of Chapter 10.

Worked Example 2.3: A Single Cross-Validated Context

Find-context Susa XII.4, associated with tablet P008832 (Case Study 1) via shared stratigraphic locus, yields a charred barley-grain sample radiocarbon dated to $4{,}380 \pm 35$ BP, calibrating (IntCal20, 2σ) to 3090–2920 BCE. The Bayesian chronological model's posterior for the same locus, derived purely from stratigraphic depth and associated glyptic style, is 3,070–2,950 BCE (95% HPD), overlapping the radiocarbon range almost completely and centering within 20 years of the radiocarbon median. This degree of agreement, replicated across the great majority of $\mathcal{D}_{14C}$, is the basis for Finding 2.4's general conclusion.

Proposition 2.3: Implications for the PE III Fragmentation Finding

Because 9 of the 34 cross-validated contexts fall within the PE III phase — proportionally over-represented relative to PE III's share of the total corpus, reflecting a deliberate sampling choice to prioritize radiocarbon validation precisely where the Bayesian chronological posteriors are least constrained by glyptic material (Section 2.1) — the strong agreement documented in Finding 2.4 extends with particular force to the PE III phase boundary central to Finding 2.1's fragmentation conclusion. This addresses a natural objection to Finding 2.1: that the apparent PE III sign-repertoire drift might instead reflect a chronological mis-assignment artificially grouping a more temporally heterogeneous set of tablets into a single nominal phase. The radiocarbon cross-validation does not support this alternative explanation, lending additional confidence to the substantive (rather than artifactual) interpretation of the PE III fragmentation pattern adopted throughout Chapters 4, 6, and 9.

© Silvano Salvador

2.6 The Proto-Elamite Network: Inter-Site Administrative Flows

The Bayesian chronological model establishes *when* tablets were produced; the transfer-entropy analysis of Chapter 4 establishes statistical *influence* directions between sites. This section complements those analyses with a structural examination of the administrative network that the tablet corpus implies, treating each site as a node and estimating inter-site administrative flow intensity from the corpus statistics.

Definition 2.6: The Inter-Site Administrative Flow Matrix F

The inter-site flow matrix $\mathbf{F} \in \mathbb{R}^{4\times4}$ is constructed for the four principal sites (Susa, Anshan, Sialk, Yahya) with entries $F_{ij} = \text{JSD}^{-1}(\hat{p}^{(i)}, \hat{p}^{(j)})$ — the reciprocal of the Jensen–Shannon divergence between the unigram frequency distributions at sites i and j, treating high similarity (low divergence) as a proxy for high administrative connectivity. The diagonal entries $F_{ii} = \infty$ (within-site similarity is trivially maximal) are excluded from all comparative analyses.

Proposition 2.4: The Network Is Susa-Centric with a Strong Anshan Secondary Hub

The off-diagonal entries of $\mathbf{F}$ reveal a star-shaped network topology centered on Susa: $F_{\text{SUS,ANS}} = 2.94$, $F_{\text{SUS,YAH}} = 2.71$, $F_{\text{SUS,SIA}} = 2.48$, $F_{\text{ANS,YAH}} = 1.82$, $F_{\text{ANS,SIA}} = 1.64$, $F_{\text{YAH,SIA}} = 1.31$. Susa has the highest mean connectivity ($\bar{F}_{\text{SUS}} = 2.71$), consistent with its role as the primary scribal innovation center (Finding 4.2). Anshan ($\bar{F}_{\text{ANS}} = 2.13$) functions as a secondary hub: more strongly connected to Susa than to Sialk or Yahya, but also internally connected to those peripheral sites more than they are to each other. Sialk and Yahya are peripheral nodes, most connected to Susa and secondarily to Anshan, with the weakest mutual connection ($F_{\text{YAH,SIA}} = 1.31$) despite their roughly comparable geographic distance from Susa. This topology is consistent with the Network Horizon framework's prediction that administrative connectivity reflects organizational links rather than simple geographic proximity.

Note 2.5: Phase-Specific Network Evolution

The inter-site flow matrix $\mathbf{F}$ computed separately for PE I, PE II, and PE III reveals a network that grows in overall connectivity from PE I to PE II (all off-diagonal F_{ij} values increase) and then contracts and fragments in PE III (off-diagonal values decrease, particularly for Sialk and Yahya connections to Susa). This phase-specific network evolution is independent confirmation, at the network-topology level, of the sign-frequency drift of Finding 2.1 and the transfer-entropy fragmentation of Note 4.1: all three methods of characterizing cross-site administrative coordination point to a peak in PE II and a disruption in PE III.

2.7 Stratigraphic Outliers and the Detection of Anomalous Provenance

The Bayesian chronological model of Definition 2.1 assigns every provenanced tablet a posterior distribution over calendar date; a small number of tablets receive posteriors that are poorly reconciled with their recorded excavation context, a pattern worth formalizing both as a chronological quality-control step and as a precursor to the authenticity analysis of Ap-

© Silvano Salvador

pendix Q.

Definition 2.7: The Stratigraphic Discordance Score

For a tablet with recorded excavation locus ℓ and posterior date distribution $\pi(\text{date} \mid \text{glyptic}, \text{morphometrics})$ (Definition 2.1, glyptic-and-morphometric terms only, excluding the locus term itself), the discordance score

$$\Delta_\ell \;=\; \big|\mathbb{E}[\text{date} \mid \pi] - \widehat{\text{date}}(\ell)\big| / \sigma_\ell$$

measures, in units of the locus-based dating's own uncertainty σ_ℓ, how far the sign- and seal-based posterior diverges from the date independently assigned to the tablet's physical findspot. This score also serves as a preliminary screen feeding into the dedicated authenticity and provenance analysis of Chapter 29.

Finding 2.5: Eleven Tablets Flagged as Stratigraphically Discordant

Of the $n = 412$ tablets with recorded stratigraphic context, 11 (2.7%) have $\Delta_\ell > 3$ (three locus-uncertainty standard deviations), a rate modestly but not dramatically above the 0.3% expected under a well-behaved Gaussian discordance distribution with no genuine outliers. Inspection of the flagged tablets' excavation records shows that 8 of the 11 come from disturbed or re-deposited loci explicitly flagged as such in the original excavation notebooks (de Mecquenem's early Susa seasons being disproportionately represented), 2 show sign-repertoire profiles consistent with the PE I/PE II phase boundary (plausibly genuine transitional tablets rather than discordant ones), and 1 remains unexplained and is flagged for renewed physical inspection.

Corollary 2.2: The Discordance Test as a Pre-Screening Filter

Because 10 of the 11 flagged tablets have an independent, archivally documented explanation for their discordance (disturbed context or genuine transitional status), Definition 2.7's test is validated as a useful pre-screening filter for stratigraphic data quality: the discordance score identifies precisely the cases that excavation-record metadata independently flags as problematic, without requiring that metadata to be consulted first. All flagged tablets are excluded from the phase-stratified analyses of Section 2.3 and elsewhere in this monograph (consistent with the $n = 387$ single-phase-assigned subset used throughout Chapter 3), a more principled exclusion criterion than the simple posterior-probability threshold used in isolation.

2.8 Propagating Phase-Assignment Uncertainty to Downstream Analyses

Every phase-stratified statistic computed elsewhere in this monograph (Proposition 3.4's diachronic entropy trend, Finding 2.1's drift rate, and others) conditions on a hard phase assignment for each tablet, even though Definition 2.3 produces a genuinely probabilistic assignment. This section quantifies the resulting understatement of uncertainty and confirms that it does not threaten the monograph's qualitative conclusions.

© Silvano Salvador

Proposition 2.5: Hard-Assignment Uncertainty Is Modest Relative to Sampling Uncertainty

Re-deriving the diachronic entropy trend of Proposition 3.4 using a fully probabilistic phase-weighting scheme (each tablet contributing to all three phase-specific entropy estimates in proportion to its posterior phase probabilities, rather than being assigned solely to its modal phase) yields PE I-to-PE III entropy increases of $+0.41$ bits (probabilistic weighting) versus $+0.45$ bits (hard assignment, as reported in Proposition 3.4): a 9% relative shift, well within the original finding's reported confidence interval. The same comparison applied to Finding 2.1's drift-rate ratio yields a 7% relative shift. Hard phase assignment, restricted as it is to the high-confidence subset (> 0.80 posterior probability), is accordingly confirmed as a defensible simplification that does not materially distort the phase-stratified findings reported throughout this monograph.

© Silvano Salvador

Chapter 3

Entropy, Zipf's Law, and the Generative Process

3.1 Shannon Entropy of the Non-Numerical Sign Sequence

Definition 3.1: The Plug-in Entropy Estimator with Miller–Madow Correction

For the non-numerical sub-corpus $\mathcal{S}$ with $N_{\mathcal{S}} = 10{,}142$ tokens over $|\Sigma^*_{PE}| = 637$ sign types, the zeroth-order Shannon entropy is estimated as

$$\hat{H}_0 = -\sum_{s \in \Sigma^*_{PE}} \hat{p}(s) \log_2 \hat{p}(s) + \frac{\hat{m} - 1}{2N_{\mathcal{S}} \ln 2},$$

where $\hat{p}(s) = n_s / N_{\mathcal{S}}$ is the empirical relative frequency, $\hat{m} = 637$ is the number of observed types, and the second term is the Miller–Madow bias correction. The first-order conditional entropy is estimated analogously, with all bigram counts subject to add-λ Laplace smoothing ($\lambda = 0.1$) to handle zero-count bigrams.

Finding 3.1: The Zeroth-Order Entropy Estimate

The Miller–Madow-corrected zeroth-order entropy of the non-numerical sub-corpus is $\hat{H}_0 = 7.21 \pm 0.09$ bits (bootstrap 95% CI, $B = 2{,}000$ resamples). This value is consistent with the theoretical maximum $\log_2(637) = 9.32$ bits at a compression ratio of 77%, indicating that the sign distribution is far from uniform — a small number of high-frequency signs dominate the corpus, and a large number of rare types form a heavy tail. For comparison, the zeroth-order entropy of the proto-cuneiform non-numerical sign corpus (Chapter 8) is $\hat{H}_0^{PC} = 6.84$ bits over a comparable effective vocabulary, and the Linear A entropy is $\hat{H}_0^{LA} = 5.22$ bits (Chapter 3 of the companion Linear A monograph). The higher proto-Elamite entropy reflects a substantially larger sign inventory at comparable corpus size.

Proposition 3.1: The First-Order Conditional Entropy

The first-order conditional entropy, estimated with Laplace-smoothed bigram counts, is $\hat{H}_1 = 5.83 \pm 0.14$ bits, yielding an entropy reduction $\Delta H_{01} = \hat{H}_0 - \hat{H}_1 = 1.38$ bits. This reduction quantifies the amount of information about the next sign conveyed by knowing the current sign: a value of 1.38 bits indicates moderate sequential depen-

dency, substantially below the $\Delta H_{01} \approx 2.0$–$2.8$ bits typical of natural-language syllabic corpora (where grammatical and phonotactic constraints produce strong next-symbol prediction), but above the $\Delta H_{01} \approx 0.2$–$0.5$ bits expected for a purely random or tallying system. This intermediate range is the first quantitative signature of the constrained-but-not-linguistic character of the corpus developed further in Section 3.3.

3.2 Zipf's Law and the Sign-Frequency Distribution

Definition 3.2: The Zipf-Mandelbrot Fit

The frequency-rank distribution of the 637 observed sign types is fitted to the Zipf-Mandelbrot law $f(r) = C(r + b)^{-\alpha}$ using maximum-likelihood estimation over the upper tail ($r \geq 10$, to avoid the head bias documented by Clauset et al. 2009 for naive rank-frequency OLS). The parameters $(\hat{\alpha}, \hat{b})$ and their bootstrap confidence intervals are reported alongside the Kolmogorov–Smirnov goodness-of-fit p-value for the fitted distribution.

Finding 3.2: A Steeper-Than-Natural-Language Zipf Slope

Maximum-likelihood fitting yields $\hat{\alpha} = 1.31$ (95% CI $[1.24, 1.39]$) and $\hat{b} = 2.1$ (95% CI $[1.4, 2.8]$). The Zipf-Mandelbrot law provides an adequate fit to the upper tail (KS $p = 0.18$). The exponent $\hat{\alpha} = 1.31$ is significantly above the range $\alpha \in [0.9, 1.1]$ typical of natural-language corpora at comparable corpus sizes (comparison against the Linear A companion monograph's reference panel of 47 natural-language corpora: proto-Elamite falls at the 97th percentile of the $\hat{\alpha}$ distribution, $p = 0.03$). A steeper Zipf slope indicates a more extreme concentration of usage on a small number of high-frequency signs, with a more rapidly attenuating long tail — consistent with the limited combinatorial range of a semasiographic commodity-coding system relative to the full vocabulary of a natural language, which requires a larger active-type vocabulary for the same number of tokens.

Worked Example 3.1: The Top-Ten Sign Types and Their Frequencies

The ten highest-frequency sign types in $\mathcal{S}$ collectively account for 31% of all non-numerical tokens. The most frequent single type (sign PE-001 in the OPEP numbering, a cross-in-circle motif associated with a cattle/livestock category in prior epigraphic literature) accounts for 4.8% of tokens alone. In a natural-language corpus of comparable size, the ten highest-frequency types would account for approximately 20%–24% of tokens (based on the reference panel of the companion monograph), and the single most frequent type for 2%–3%. The higher proto-Elamite concentration is consistent with Definition 1.6's prediction that a semasiographic system's most frequent signs — universal commodity classifiers, common institutional designators — bear a disproportionate share of the token load relative to the more uniformly distributed closed-class vocabulary of a natural language.

© Silvano Salvador

3.3 The Positional Entropy Profile and the Restricted-Register Anomaly

Definition 3.3: The k-th Order Conditional Entropy Profile

The k-th order conditional entropy H_k is defined as the Shannon entropy of the sign at position $k+1$ conditional on the signs at positions 1 through k within a case (sign-group bounded by incised rulings), computed over all cases of length at least $k+1$ in the corpus. In natural-language corpora, H_k is typically monotonically decreasing in k, approaching a residual value $H_\infty > 0$ determined by the irreducible entropy of the language at the given register and formality level.

Finding 3.3: The Restricted-Register Anomaly

Computing H_k for $k = 0, 1, 2, 3$ using regularized estimation (see Appendix B):

$$H_0 = 7.21, \quad H_1 = 5.83, \quad H_2 = 5.61, \quad H_3 = 5.58 \text{ bits.}$$

The conditional entropy profile shows a large drop from H_0 to H_1 (-1.38 bits), a much smaller drop from H_1 to H_2 (-0.22 bits), and a near-zero drop from H_2 to H_3 (-0.03 bits, within bootstrap estimation uncertainty). The profile thus *plateaus* after first order at approximately 5.6 bits — a value this monograph terms the *restricted-register anomaly*. Natural-language corpora of comparable size show a continuing, if decelerating, decline through at least $k = 3$, with H_3 typically 0.5–1.2 bits below H_1. The proto-Elamite plateau at $k = 1$ is inconsistent with the sequential dependency structure of any tested natural-language corpus at significance level $p < 0.01$ (permutation test against the companion monograph's reference panel), and is the primary information- theoretic evidence against $\mathcal{H}_1$ developed in this monograph.

Note 3.1: Alternative Explanations of the Plateau

Three alternative explanations of the early conditional entropy plateau are considered: (i) *Corpus-size artefact*: the plateau could reflect insufficient corpus size for reliable second-order estimation. This is addressed by comparing against synthetic corpora of the same size drawn from natural-language models: none produces a plateau as flat as the proto-Elamite profile at this corpus size ($p < 0.01$). (ii) *Case-length truncation*: if cases are short, the entropy profile is computed over few within-case positions. The result holds when restricted to long cases (≥ 5 signs), so truncation does not account for it. (iii) *Heavy tokenization noise*: erroneous sign-boundary identification could flatten the entropy profile artificially. Restricting to the best-quality OPEP scans (curators' confidence rating $\geq 4/5$) does not change the plateau, ruling out this explanation.

Corollary 3.1: The Anomaly Quantified as an MDL Differential

Under the MDL framework of Chapter 7, the restricted-register anomaly corresponds to a specific compressive advantage of $\mathcal{H}_2$ over $\mathcal{H}_1$: a first-order Markov model captures nearly all the sequential structure present in the corpus, while the additional parameters of a higher-order natural-language model provide negligible compression gain. The MDL differential attributable to this finding alone (computed in Chapter 7) is $\Delta\text{MDL}_{12} = 1{,}840$ bits in favour of $\mathcal{H}_2$, corresponding to a Bayes factor of approximately 10^{554} against $\mathcal{H}_1$ on this evidence stream alone — an extreme value driven by

© Silvano Salvador

the large corpus size that amplifies even a small per-token structural difference into a large model-comparison statistic.

3.4 Numerical versus Non-Numerical Sub-System Entropy Contrast

Proposition 3.2: The Numerical Sub-System Entropy Is Lower

The zeroth-order entropy of the numerical sub-corpus $\mathcal{N}$ (treating each numerical notation token as a sign type from its own inventory) is $\hat{H}_0^{\mathcal{N}} = 4.31$ bits, substantially lower than $\hat{H}_0^{\mathcal{S}} = 7.21$ bits for the non-numerical sub-corpus. The conditional entropy profile of $\mathcal{N}$ drops sharply to $H_1^{\mathcal{N}} = 2.14$ bits and plateaus immediately ($H_2^{\mathcal{N}} = 2.10$ bits), consistent with a positional-value numerical notation in which each token's value is largely determined by its position within the notation and the numeral preceding it — exactly what is expected for a well-understood metrological system with a small number of systematic notation conventions.

Finding 3.4: The Internal-Control Value of the Dual Sub-System

The contrast between the numerical and non-numerical entropy profiles provides an internal control absent from the Voynich corpus (which has no independently understood sub-system) and from the Linear A corpus (whose numerical sub-system is understood but its sign types are not fully separable from the logographic sub-system in the existing transcriptions). Proto-Elamite is thus uniquely positioned, among the three undeciphered scripts treated in this series, to demonstrate directly, within the same corpus, that the statistical methodology produces qualitatively different profiles for a known notation (the numerical system) and an unknown system (the non-numerical signs). This internal validation strengthens confidence that the restricted-register anomaly of Finding 3.3 reflects a genuine property of the non-numerical sign system rather than a methodological artefact.

3.5 Scribal-Hand Variation and the Robustness of the Entropy Profile

Note 3.2: Motivation

The entropy and Zipf statistics of Sections 3.1–3.3 are computed over the non-numerical sub-corpus $\mathcal{S}$ pooled across all identified scribal hands. This section asks whether the pooled profile, and in particular the restricted-register anomaly central to this monograph's argument, is a genuine corpus-wide property or instead an artifact of pooling together hands with substantially different individual entropy profiles — a concern directly analogous to the well-known statistical phenomenon in which pooling heterogeneous sub-populations can manufacture spurious aggregate patterns (Simpson's paradox-adjacent reasoning).

© Silvano Salvador

Definition 3.6: Hand-Stratified Entropy Estimation

Using the morphometric hand-identification results of Section 2.4 (Proposition 2.2), the 19 scribal hands identified with confidence ≥ 0.8 and contributing at least 200 non-numerical sign tokens each (a coverage threshold required for stable individual entropy estimation) are isolated, and the entropy plateau ratio H_∞/H_0 and Zipf exponent α_Z are independently re-estimated within each hand's sub-corpus, using the same Miller–Madow bias-corrected estimators as the pooled analysis (Appendix B.1).

Finding 3.5: The Restricted-Register Profile Is Stable Across Scribal Hands

The 19 hand-specific entropy plateau ratios range from 0.58 to 0.74, with a mean of 0.65 and standard deviation 0.045 — closely centered on, and statistically indistinguishable from (one-sample t-test against the pooled value 0.66, $t = 0.31$, $p = 0.76$), the pooled corpus estimate of Finding 3.3. Critically, *no* individual hand's entropy plateau ratio falls within the natural-language range of $[0.25, 0.45]$ established for the syllabic comparanda in Chapter 5, and the between-hand variance ($0.045^2 \approx 0.002$) is small relative to the gap separating the observed range from that natural-language benchmark. This result directly addresses the pooling concern of Note 3.2: the restricted-register anomaly is not an artifact of aggregating linguistically diverse individual hands into a single profile, but is independently reproduced within essentially every individual scribal hand large enough to support separate estimation.

Worked Example 3.2: The Two Most Divergent Hands

The two hands with the most extreme individual entropy plateau ratios are Hand H-07 ($H_\infty/H_0 = 0.58$, the lowest in the sample) and Hand H-14 ($H_\infty/H_0 = 0.74$, the highest). Both hands are independently identified, via the stroke-morphometric clustering of Section 2.4, as PE II-phase Susa hands of broadly comparable tablet-count productivity, ruling out an obvious confound of phase or site. Inspection of Hand H-14's sign usage shows a higher proportion of tokens drawn from the high-frequency sign tail (consistent with this hand specializing in a narrower range of transaction types), offering one plausible, non-linguistic explanation for its comparatively higher entropy plateau — itself a finding consistent with, rather than contrary to, the H2 (constrained semasiography) interpretation favored in Chapter 7, since H2 predicts exactly this kind of scribe-to-scribe variation in deployed sub-vocabulary breadth without requiring any corresponding variation in linguistic competence.

Proposition 3.3: An Upper Bound on Hand-Level Heterogeneity Effects

Treating the between-hand variance in H_∞/H_0 as an upper bound on any hand-specific "registers" that might exist within a hypothetical underlying natural language (the H1′ scenario of Note 6.2), the observed between-hand standard deviation of 0.045 implies that even a maximally heterogeneous mixture of hand-specific linguistic registers could not, on its own, account for the $\approx$ 0.2–0.3 gap between the observed pooled plateau ratio (0.66) and the natural-language benchmark range ($[0.25, 0.45]$) established in Chapter 5. This strengthens, on a hand-stratified statistical basis independent of the classification-space and Bayesian-synthesis arguments of Chapters 5 and 7, the case against an unrestricted natural-language account of the non-numerical sign

© Silvano Salvador

system.

3.6 Rényi Entropy Generalization and the Sign-Concentration Profile

Shannon entropy $H_0 = H_1(\text{Shannon})$ is one member of the Rényi entropy family H_α for $\alpha \geq 0$, $\alpha \neq 1$. The Rényi family provides a richer characterization of sign-frequency concentration than the Shannon estimate alone: H_α for $\alpha < 1$ is sensitive to the rare-sign tail, while H_α for $\alpha > 1$ is dominated by the most frequent signs. Mapping the proto-Elamite sign frequency distribution across the Rényi parameter α produces a concentration profile that characterizes the distributional shape beyond what a single entropy value captures.

Definition 3.7: The Rényi Entropy Profile

For the non-numerical sub-corpus $\mathcal{S}$ with empirical frequency distribution $\hat{p}$, the order-α Rényi entropy is

$$H_\alpha(\hat{p}) = \frac{1}{1-\alpha} \log_2 \left(\sum_{s \in \Sigma^*_{PE}} \hat{p}(s)^\alpha \right), \quad \alpha \geq 0,\ \alpha \neq 1,$$

with $H_1(\hat{p}) = \lim_{\alpha \to 1} H_\alpha(\hat{p}) = H_{\text{Shannon}}(\hat{p})$ (the standard Shannon entropy). The Rényi profile is the function $\alpha \mapsto H_\alpha(\hat{p})$ evaluated at $\alpha \in \{0.25, 0.5, 1.0, 2.0, 3.0, 5.0\}$.

Finding 3.6: The Proto-Elamite Rényi Profile Confirms Heavy-Tailed Concentration

The Rényi profile for $\mathcal{S}$:

α	0.25	0.50	1.0	2.0	3.0	5.0
H_α (bits)	9.17	8.69	7.21	5.42	4.81	4.22

The steep descent from $H_{0.25} = 9.17$ to $H_5 = 4.22$ bits quantifies the extreme concentration of token mass on a small number of high-frequency signs: the Rényi-2 (collision entropy) value of 5.42 bits implies an effective vocabulary size of $2^{5.42} \approx 43$ signs at the "collision" level, despite a nominal inventory of 637 types. For comparison, the Linear B reference corpus ($\mathcal{R}_3$) has a Rényi-2 entropy of 3.71 bits (effective vocabulary ≈ 13 signs), reflecting its much smaller syllabic inventory; the synthetic constrained-semasiography control ($\mathcal{R}_6$) has Rényi-2 entropy of 5.18 bits (effective vocabulary ≈ 36 signs), the closest match to proto-Elamite among all reference systems and consistent with the H2-favoring classification of Chapter 5.

3.7 Diachronic Entropy Shift: Testing for Temporal Trends

The three-phase chronological model of Chapter 2 enables a diachronic test that is not possible from the corpus aggregate alone: whether the entropy profile of the non-numerical sub-corpus changes systematically across PE I, PE II, and PE III. A significant diachronic entropy shift would bear on the fragmentation hypothesis of Finding 2.1 by providing an information-theoretic parallel to the sign-frequency-divergence result.

© Silvano Salvador

Proposition 3.4: A Diachronic Entropy Trend Is Detectable at the Phase Level

Computing the Miller–Madow-corrected Shannon entropy $\hat{H}_0$ and the entropy plateau ratio H_∞/H_0 separately for the three phase-assigned sub-corpora (restricting to tablets with phase posterior probability > 0.80):

Phase	N (tokens)	$\hat{H}_0$ (bits)	H_∞/H_0
PE I	$\approx 2{,}100$	6.94 ± 0.14	0.68 ± 0.06
PE II	$\approx 5{,}800$	7.18 ± 0.09	0.65 ± 0.04
PE III	$\approx 2{,}200$	7.39 ± 0.13	0.69 ± 0.06

The zeroth-order entropy increases monotonically from PE I to PE III ($+0.45$ bits total, significant: t-test PE I vs. PE III, $t = 3.1$, $p = 0.002$), consistent with an expanding and diversifying sign repertoire over time. The entropy plateau ratio is non-monotonic: lower in PE II than in PE I or PE III. This non-monotonicity is interpretable: PE II represents the period of maximum administrative standardization (maximum slot-grammar regularity), while PE I shows looser standardization (Case Study 2) and PE III shows fragmentation-driven loosening (Finding 2.1 and Case Study 8). The diachronic entropy trend thus corroborates the chronological model of Chapter 2 from an entirely independent (information-theoretic) angle.

3.8 Mutual Information Between the Numerical and Non-Numerical Sub-Systems

Proposition 3.2 established that the numerical sub-system $\mathcal{N}$ exhibits substantially lower marginal entropy than the non-numerical sub-system $\mathcal{S}$, but a marginal-entropy contrast alone is silent on whether the two sub-systems are *statistically coupled* within a single tablet: a quantity, a commodity designation, and an administrative formula could in principle co-vary tightly (high mutual information, suggesting a tightly integrated notation) or could behave as near-independent channels glued together only by physical co-location on the clay surface (low mutual information, suggesting two loosely coupled sub-grammars). This section formalizes and estimates that coupling.

Definition 3.8: Sub-System Mutual Information

For a tablet-line ℓ jointly emitting a numerical token class $\nu(\ell) \in \mathcal{V}_\nu$ (one of the ≈ 6 coarse magnitude bins of the dominant notation system in use on that line) and a non-numerical sign class $\sigma(\ell) \in \mathcal{V}_\sigma$ (the line-initial commodity or office sign, drawn from the ≈ 40 most frequent such signs), the mutual information between the two channels is

$$I(\nu;\sigma) \;=\; \sum_{v \in \mathcal{V}_\nu} \sum_{s \in \mathcal{V}_\sigma} p(v,s) \log_2 \frac{p(v,s)}{p(v)\,p(s)},$$

estimated from the joint empirical distribution over $n = 8{,}214$ lines bearing both a classifiable numerical entry and a classifiable line-initial sign, with the standard Miller–Madow bias correction applied to the joint, marginal, and conditional entropy terms entering the estimator.

© Silvano Salvador

Finding 3.7: Moderate but Highly Significant Cross-Channel Coupling

The estimated mutual information is $\hat{I}(\nu;\sigma) = 0.71 \pm 0.05$ bits, small relative to either marginal entropy ($\hat{H}(\nu) \approx 2.1$ bits, $\hat{H}(\sigma) \approx 4.9$ bits) but far in excess of the permutation-null expectation $I_0 = 0.00 \pm 0.02$ bits obtained by independently shuffling the numerical and non-numerical channels across lines within each tablet (10,000 permutations; $z = 34.7$). The normalized coupling coefficient $U(\nu;\sigma) = I(\nu;\sigma)/H(\sigma) \approx 0.145$ indicates that knowledge of the commodity sign reduces uncertainty about the numerical notation system in use by roughly 14.5% — consistent with the well-attested practice (Friberg, 1978; Englund, 2004) whereby specific commodities are conventionally counted in specific notation systems (grain in system S, discrete objects in system B, and so forth), but far short of the coupling that would be expected if commodity identity rigidly determined notation choice ($U \to 1$).

Corollary 3.2: An Upper Bound on Purely Compositional Readings

Because $U(\nu;\sigma) \ll 1$, any decipherment hypothesis in which the non-numerical sign sequence functions as a fully determined morphological extension of the numerical notation (e.g. a hypothesis under which commodity signs are mere classifiers mechanically selected by the counting system) is disfavoured: such a hypothesis predicts $U \to 1$ in the limit of a finite classifier inventory, two orders of magnitude above the observed value. The finding is consistent with, and quantitatively sharpens, the qualitative observation of Dahl (2005) that proto-Elamite commodity signs retain independent combinatorial freedom.

3.9 Compound Sign-Groups and the Boundary-Inference Problem

Unlike Linear A, the proto-Elamite non-numerical sub-corpus possesses no explicit word-divider mark; segmentation into multi-sign administrative units must instead be inferred from line breaks, case (sealed compartment) boundaries on the tablet surface, and recurring co-occurrence structure. This section develops a formal boundary-inference statistic and applies it to characterize the distribution of inferred sign-group lengths.

Definition 3.9: The Boundary Strength Statistic

For adjacent sign positions $i, i+1$ within a line, define the boundary strength

$$B_i \;=\; \log_2 \frac{p(\text{sign}_i, \text{sign}_{i+1})}{p(\text{sign}_i)\,p(\text{sign}_{i+1})} \Big/ -1,$$

the negated pointwise mutual information of the adjacent pair. A position i is classified as a candidate group boundary when B_i exceeds the 70th-percentile threshold of the pooled B_i distribution and is simultaneously not contradicted by the case-boundary evidence (Section 1.7 of the metrological appendix); this combines a purely distributional signal with the physical layout of the tablet.

© Silvano Salvador

Finding 3.8: A Right-Skewed, Approximately Log-Normal Group-Length Distribution

Applying Definition 3.9 across the full non-numerical sub-corpus yields $n = 4{,}871$ inferred sign-groups with a mean length of 2.34 signs and a strongly right-skewed distribution (skewness $= 1.87$) that is well fit by a log-normal distribution ($\mu_{\log} = 0.71$, $\sigma_{\log} = 0.52$; Kolmogorov–Smirnov $D = 0.031$, $p = 0.21$, not rejected) rather than by a geometric distribution ($D = 0.089$, $p < 0.001$, rejected) of the kind expected under a memoryless sign-emission process. The rejection of the geometric null is itself informative: a memoryless (zeroth-order Markov) generative process would necessarily produce geometrically distributed group lengths under any boundary-marking convention, so the observed log-normal shape constitutes independent evidence — convergent with the n-gram findings of Chapter 4 — for a generative process with internal structure extending beyond single-step dependence.

Note 3.3: Comparison with the Linear A Word-Length Distribution

The proto-Elamite group-length distribution (mean 2.34 signs, log-normal shape) is shorter on average and less variable than the Linear A word-length distribution derived from the explicit word-divider evidence (mean ≈ 3.1 signs; companion volume, Chapter 3, Section 5), a difference plausibly attributable to the more restricted, formulaic character of proto-Elamite administrative notation relative to whatever degree of natural-language encoding underlies Linear A. The comparison should be treated cautiously, however, since the proto-Elamite figure is itself an inferred quantity subject to the boundary-detection methodology of Definition 3.9, whereas the Linear A figure rests on an explicit orthographic convention.

3.10 Cross-Entropy Against Frequency-Preserving Synthetic Nulls

The information-theoretic case against a pure tallying or single-symbol counting interpretation of the non-numerical sub-system (Abstract; Finding 3.1) rests in part on comparison against synthetic null corpora. This section makes that comparison fully explicit and reports the cross-entropy gap on which it depends.

Definition 3.10: The Frequency-Preserving Shuffle Null

A synthetic corpus is constructed by drawing, independently for each tablet line, sign tokens with replacement from the empirical unigram frequency distribution $\hat{p}(\text{sign})$ of the real corpus, holding the number of tokens per line fixed at its observed value. This null preserves the marginal (zeroth-order) sign frequencies exactly while destroying all higher-order sequential structure, and is distinct from the bootstrap resampling procedure used elsewhere in this monograph for confidence-interval construction.

Finding 3.9: A Large and Significant Cross-Entropy Gap

The cross-entropy of the real corpus under a second-order Markov model trained on the synthetic null is $H_{\times} = 8.94$ bits/sign, against a within-sample second-order entropy of $\hat{H}_2 = 6.81$ bits/sign for the real corpus itself (Section 3.3): a gap of 2.13 bits/sign, or a 24% relative reduction in per-sign uncertainty attributable to genuine sequential struc-

© Silvano Salvador

ture absent from the shuffle null. Averaged over 200 independent synthetic-null draws, the gap is stable to within ± 0.04 bits, and the real corpus's within-sample entropy lies 46 synthetic-null standard deviations below the null-model mean cross-entropy — a margin that, combined with the n-gram divergence measures of Chapter 4, leaves negligible posterior mass on any hypothesis under which the non-numerical sign sequence is generated by a zeroth-order (frequency-only) process.

© Silvano Salvador

Chapter 4

n-gram Models, Transition Matrices, and Cross-Entropy Comparisons

4.1 Markov Models of the Proto-Elamite Sign Sequence

Definition 4.1: The Transition Matrix T

The first-order Markov transition matrix $\mathbf{T} \in \mathbb{R}^{637\times637}$ is estimated from within-case consecutive sign-pair counts: $\hat{T}_{ij} = (c_{ij} + \lambda)/(\sum_k c_{ik} + \lambda|\Sigma^*_{PE}|)$, where c_{ij} is the number of times sign j follows sign i within a case, and $\lambda = 0.1$ is the Laplace smoothing parameter. The matrix is sparse: 91.3% of entries are zero before smoothing (i.e. the corresponding bigram is unattested), reflecting the restricted combinatorics of the non-numerical sign sequence. For comparison, the Linear A companion monograph reports a bigram-zero proportion of 67% for the Linear A corpus at comparable vocabulary size, consistent with the richer sequential dependency structure expected for a natural-language corpus.

Proposition 4.1: The Perplexity of the First-Order Model

The per-token perplexity of the smoothed first-order Markov model, evaluated on a held-out 20% random split of the non-numerical corpus, is $\mathrm{PP}_1 = 43.7$ (95% CI $[41.1, 46.3]$ by bootstrap). This is compared against:

- A zeroth-order (unigram) baseline: $\mathrm{PP}_0 = 75.4$;
- A second-order (trigram) model: $\mathrm{PP}_2 = 42.9$ (95% CI $[40.2, 45.7]$), not significantly lower than PP_1 ($p = 0.41$, bootstrap overlap test), confirming the entropy-profile plateau of Chapter 3 in a held-out-prediction framework;
- The proto-cuneiform first-order perplexity (Section 4.2): $\mathrm{PP}_1^{PC} = 38.1$, moderately lower, consistent with the proto-cuneiform system's somewhat smaller sign inventory and higher within-case sequential regularity.

Worked Example 4.1: Computing Perplexity on a Single Case

A representative five-sign case from a Susa PE II tablet (CDLI P008832, case 3: *PE-001 PE-047 PE-001 PE-312 PE-001*) yields a first-order Markov log-probability $\log_2 \hat{p}(\text{case}) = \log_2 \hat{p}(\text{PE-001}) + \sum_{i=1}^{4} \log_2 \hat{T}_{s_i,s_{i+1}} = -7.21 + (-5.83 - 7.14 - 4.91 - 5.83) = -30.92$ bits, corresponding to a per-token perplexity of $2^{30.92/5} = 2^{6.18} = 72.5$

for this specific case — above the corpus mean of 43.7, reflecting the high-entropy PE-312 transition in position 3–4. The case illustrates the common pattern of a high-frequency anchor sign (PE-001) framing lower-frequency content signs in medial positions.

4.2 Cross-Entropy Against the Proto-Cuneiform Reference Corpus

Definition 4.2: The Cross-Entropy Test

For two corpora A and B, the cross-entropy of B under a model trained on A is $H(B\|\hat{p}_A) = -\frac{1}{N_B}\sum_i \log_2 \hat{p}_A(s_i^B)$, where the sum is over all tokens s_i^B in corpus B. A low cross-entropy indicates that the statistical model of A predicts the tokens of B well, suggesting that the two corpora share a common generative structure; a high cross-entropy indicates divergence. The cross-entropy test is applied symmetrically: $H(\text{PE}\|\hat{p}_{PC})$ and $H(\text{PC}\|\hat{p}_{PE})$ are both computed.

Finding 4.1: Proto-Elamite and Proto-Cuneiform Are Mutually Poorly Predicted

The cross-entropy of the proto-Elamite corpus under a unigram model trained on proto-cuneiform is $H(\text{PE}\|\hat{p}_{PC}) = 9.14$ bits, substantially higher than the within-corpus perplexity of 7.21 bits, indicating that the proto-cuneiform unigram model is a poor predictor of proto-Elamite sign choices. The reverse cross-entropy $H(\text{PC}\|\hat{p}_{PE}) = 8.83$ bits is similarly elevated relative to the proto-cuneiform within-corpus entropy of 6.84 bits. The symmetric KL divergence $D_{KL}(P_{PE}\|P_{PC}) + D_{KL}(P_{PC}\|P_{PE}) = 2.30 + 1.99 = 4.29$ bits is large, indicating that the two sign frequency distributions are substantially different despite the two systems' shared administrative context, physical medium, and approximate chronological contemporaneity. This finding is inconsistent with a hypothesis of direct lexical borrowing between the two script traditions and consistent with independent sign-system development, though it does not rule out a shared conceptual framework underlying independently-chosen graphic conventions.

4.3 Higher-Order Models and the Curse of Sparsity

Proposition 4.2: The Effective Model-Order Ceiling

The effective model-order ceiling k^* — the highest k for which a k-gram model achieves statistically significant held-out perplexity improvement over a $(k-1)$-gram model on the proto-Elamite corpus — is $k^* = 1$, established by comparing bootstrap confidence intervals for PP_k at successive orders. For $k \geq 2$, the perplexity improvement $(\text{PP}_{k-1} - \text{PP}_k)$ is not statistically significant at $\alpha = 0.05$ after Holm–Bonferroni correction, reflecting the curse of dimensionality: at corpus size $N_S \approx 10{,}000$, the $637^2 \approx 406{,}000$ possible bigram types already far exceed the number of observed bigram tokens ($\approx 9{,}500$, excluding end-of-case positions), and the $637^3 \approx 258{,}000{,}000$ possible trigram types make unregularised trigram estimation entirely infeasible.

© Silvano Salvador

Definition 4.3: The Kneser–Ney Back-Off Strategy

For the Bayesian model comparison of Chapter 7, which requires higher-order probability estimates, a modified Kneser–Ney smoothing scheme is applied: the probability of a sign s following context c_{k-1} is estimated as

$$p_{KN}(s \mid c_{k-1}) = \frac{\max(c(c_{k-1}, s) - d, 0)}{\sum_{s'} c(c_{k-1}, s')} + \lambda(c_{k-1})\, p_{KN}(s \mid c_{k-2}),$$

where $d = 0.75$ is the standard Kneser–Ney discount, $\lambda(c_{k-1})$ is the interpolation weight, and the back-off recurses to lower orders. This estimator is selected over alternatives (Witten–Bell, Laplace) because of its superior performance on held-out data in sparse-vocabulary conditions, as documented in the benchmark of Chen and Goodman (1998) and confirmed on the proto-Elamite data.

4.4 Transfer Entropy Across Site and Phase Boundaries

Definition 4.4: Transfer Entropy

The transfer entropy from process Y (sign sequence at site or phase Y) to process X (sign sequence at site or phase X) is

$$T_{Y \to X} = \sum_{x_{t+1}, x_t^{(k)}, y_t^{(l)}} p(x_{t+1}, x_t^{(k)}, y_t^{(l)}) \log_2 \frac{p(x_{t+1} \mid x_t^{(k)}, y_t^{(l)})}{p(x_{t+1} \mid x_t^{(k)})},$$

where $x_t^{(k)}$ and $y_t^{(l)}$ are length-k and length-l history vectors. At $k = l = 1$, this reduces to a comparison between the transition probabilities with and without knowledge of the other site's sign at the same corpus position, operationalized here by treating same-commodity-class tablet pairs from different sites as conceptually concurrent records and computing the conditional mutual information between their sign choices.

Finding 4.2: Asymmetric Transfer Entropy Favours Susa as the Origin Site

Computing $T_{Susa \to Anshan}$ and $T_{Anshan \to Susa}$ for commodity-sign choices (restricting to the four commodity categories with at least 50 tablets at both sites in the PE II phase), the directional asymmetry $\Delta T = T_{Susa \to Anshan} - T_{Anshan \to Susa} = 0.043$ bits (95% bootstrap CI $[0.011, 0.076]$ bits, significantly positive at $p = 0.01$) indicates that knowledge of Susa's sign choices for a given commodity category marginally but significantly improves prediction of Anshan's choices, while the reverse is not true. This asymmetry is consistent with Susa functioning as the primary scribal innovation site from which sign conventions diffused outward to Anshan and Sialk — consistent with Susa's substantially larger corpus and earlier phase-I attestations — though the magnitude of the asymmetry is small and the conclusion is accordingly tentative at the current corpus size.

© Silvano Salvador

Note 4.1: Transfer Entropy Across Phase Boundaries

Applying the same framework across phase boundaries (PE I-to-PE II and PE II-to-PE III), the PE I-to-PE II transfer entropy is positive and significant ($\Delta T = 0.061$ bits, $p = 0.008$), consistent with PE II inheriting its sign conventions from PE I rather than reinventing them independently. The PE II-to-PE III transfer entropy is lower ($\Delta T = 0.019$ bits, $p = 0.11$, not significant at $\alpha = 0.05$), consistent with the sign-repertoire drift and fragmentation identified in Finding 2.1: as PE III sites diverge from the PE II standard, the directional influence of PE II on PE III becomes weaker and statistically undetectable at the available corpus size.

4.5 Benchmarking Against the Synthetic Semasiography Control

Note 4.2: Why a Synthetic Benchmark Is Needed Here

The cross-entropy comparison against proto-cuneiform (Finding 4.1) tests proto-Elamite against a real, independently attested administrative system. This section complements that test with a comparison against the synthetic constrained-semasiography control $\mathcal{R}_6$ introduced in Definition 5.1 (anticipated here, ahead of its full introduction in Chapter 5, because the relevant comparison is fundamentally an n-gram/cross-entropy exercise belonging methodologically with the present chapter). Because $\mathcal{R}_6$'s generative grammar is known by construction, this comparison provides a direct test of how well the n-gram and transfer-entropy machinery of this chapter recovers known ground-truth structure, complementing the purely comparative proto-cuneiform test of Section 4.2.

Definition 4.5: The Synthetic Semasiography Generative Grammar

The synthetic control $\mathcal{R}_6$ (full specification in Appendix B, generation parameters) is generated from a hand-built context-free slot grammar with three obligatory slots (commodity, quantity-qualifier, institution) and one optional slot (personnel identifier), each populated from a fixed per-slot sub-vocabulary of 25–40 symbols, with slot order fixed and symbol selection within each slot governed by a Zipfian sampling distribution ($\alpha_{\text{synth}} = 0.80$, chosen to match the empirically observed proto-Elamite Zipf exponent of Finding 3.2 as a deliberate design choice enabling like-for-like comparison).

Finding 4.3: The n-gram Machinery Recovers the Known Synthetic Structure

Applying the full n-gram and transition-matrix apparatus of Sections 4.1–4.3 to $\mathcal{R}_6$ recovers, as expected given the known generative process, an effective model-order ceiling of $k^*_{\mathcal{R}_6} = 1$ (matching the known slot-grammar's first-order Markov sufficiency by construction) and a positional regularity $\rho_{pos} = 0.41$ (Finding 5.1), closely matching the fixed four-slot template. The cross-entropy between $\mathcal{R}_6$ and the proto-Elamite corpus, $H(\mathcal{S}\|\hat{p}_{\mathcal{R}_6}) = 7.58$ bits, is substantially lower than the proto-Elamite/proto-cuneiform cross-entropy of 9.14 bits (Finding 4.1) and close to proto-Elamite's own within-corpus entropy of 7.21 bits, indicating that a model trained on the synthetic slot grammar predicts proto-Elamite sign sequences nearly as well as proto-Elamite predicts itself — the strongest single piece of n-gram-level evidence, anywhere in this

© Silvano Salvador

monograph, in favor of the slot-grammar account central to H2.

Worked Example 4.2: Why This Result Is Not Circular

A natural concern is that Finding 4.3's result is circular, since $\mathcal{R}_6$'s Zipf exponent was deliberately tuned to match proto-Elamite (Definition 4.5). This concern is addressed directly: the Zipf exponent alone constrains only the unigram (zeroth-order) frequency distribution, whereas the cross-entropy advantage of Finding 4.3 is computed at the bigram level and is therefore sensitive to $\mathcal{R}_6$'s *sequential* slot-grammar structure, which was fixed independently of any proto-Elamite sequential statistic. As a direct check, a second synthetic control $\mathcal{R}_6'$ is generated with the same tuned Zipf exponent but with symbol order randomly shuffled within each case (destroying the slot-grammar sequential structure while preserving the unigram frequency distribution exactly). The cross-entropy between proto-Elamite and $\mathcal{R}_6'$ is $H(\mathcal{S}\|\hat{p}_{\mathcal{R}_6'}) = 8.81$ bits, far closer to the proto-cuneiform benchmark than to $\mathcal{R}_6$'s 7.58 bits, confirming that Finding 4.3's result depends specifically on the slot-grammar sequential structure and not merely on the matched marginal frequency distribution.

4.6 Variable-Order Markov Modeling via Context-Tree Weighting

The fixed-order n-gram models of Sections 4.1–4.3 face an inherent order/sparsity trade-off (Proposition 4.2): low-order models are well estimated but coarse, high-order models are fine-grained but data-starved. A variable-order approach, in which the effective context length is allowed to differ sign by sign according to how much predictive information that context actually carries, offers a principled resolution and is adopted here via the context-tree weighting (CTW) algorithm (Willems, Shtarkov, and Tjalkens, 1995).

Definition 4.6: Context-Tree Weighted Probability

For a maximum context depth D, the CTW estimator recursively mixes, at every node of the suffix context tree, a fixed-order Krichevsky–Trofimov estimate at that node with the weighted estimate of its two children, producing a sign-sequence probability $P_{\text{CTW}}^{(D)}$ that is provably within a bounded redundancy of the best fixed-order Markov model of any order $\leq D$, without requiring that order to be specified in advance.

Finding 4.4: The CTW Model Selects an Effective Order Near Three

Fitting the CTW estimator with $D = 6$ to the proto-Elamite non-numerical sub-corpus and inspecting the posterior weight assigned to each context depth at the tree root shows that depths 2 and 3 jointly carry 81% of the total posterior weight, with depth ≥ 4 contributing a combined 6%. The CTW compression rate, 1.94 bits/sign, improves on the best fixed-order model (second-order, $\hat{H}_2 = 6.81$ bits/sign per Section 3.3 once converted to a comparable per-sign units — the CTW figure reported here uses a binary-encoded sign representation, see Appendix B) by a margin that is modest but consistent across five-fold cross-validation, confirming that little additional sequential structure remains to be captured beyond order 3, and that the second/third-order models used elsewhere in this monograph are not artificially truncating a substantially longer-range dependency structure.

© Silvano Salvador

4.7 Perplexity Comparison Across the Four Principal Hypotheses

Sections 4.1–4.6 develop n-gram machinery in service of specific local comparisons (against proto-cuneiform, against synthetic controls); this section consolidates those results into a single comparative table evaluated against the two generative hypotheses, H1 and H2, that admit a direct whole-corpus sequence-model formalization (Chapter 6 also evaluates H3 and H4, but these concern, respectively, a minority sign subset and a cross-site standardization pattern rather than a single global generative grammar, and so are not commensurable with a single perplexity figure). This provides a bridge between the sequence-level statistics of this chapter and the hypothesis-level Bayesian synthesis of Chapter 7.

Proposition 4.3: A Consolidated Cross-Hypothesis Perplexity Table

Held-out perplexity (lower is better; 10-fold cross-validation, second-order models, held-out token count $\approx 2{,}400$ per fold) is as follows:

Generative model	Held-out perplexity	Δ vs. empirical
Pure tally/counting null (excluded a priori, Finding 3.1)	187.4	+131.9
H1: Full natural-language (syllabic) encoding, $\mathcal{R}_2/\mathcal{R}_3$-trained	61.9	+6.4
H2: Constrained slot-grammar semasiography, $\mathcal{R}_6$-trained	58.6	+3.1
Empirical (within-corpus, 2nd-order)	55.5	—

The pure-tally null is decisively excluded by perplexity alone (over three times the empirical value), consistent with Finding 3.9's cross-entropy gap. Between the two combinatorially rich global models, the H2-aligned slot-grammar account ($\mathcal{R}_6$-trained, as in Finding 4.3) achieves the lower held-out perplexity and the smaller gap to the empirical ceiling, foreshadowing — but not by itself determining, since perplexity does not penalize model complexity — the MDL ranking formally derived in Chapter 7.

4.8 Skip-Gram Statistics and Non-Adjacent Co-occurrence

Standard n-gram models capture only contiguous dependency; administrative formulae, however, plausibly encode dependencies between non-adjacent slot positions (for instance, between a line-initial commodity sign and a line-final office or personnel sign, with an intervening numerical entry). This section tests for such structure directly.

Definition 4.7: The k-Skip Bigram Statistic

For skip distance $k \in \{1, 2, 3\}$, the k-skip pointwise mutual information between sign positions i and $i+k+1$ (i.e. separated by exactly k intervening tokens) is estimated as

$$\mathrm{PMI}_k(a, b) \;=\; \log_2 \frac{p(\mathrm{sign}_i = a,\ \mathrm{sign}_{i+k+1} = b)}{p(\mathrm{sign}_i = a)\, p(\mathrm{sign}_{i+k+1} = b)},$$

averaged over all sign pairs (a, b) weighted by their joint frequency to produce a single skip-distance-specific summary statistic $\overline{\mathrm{PMI}}_k$.

© Silvano Salvador

Finding 4.5: Non-Adjacent Structure Decays but Does Not Vanish

The weighted average skip-PMI is $\overline{\text{PMI}}_1 = 0.34$ bits, $\overline{\text{PMI}}_2 = 0.19$ bits, and $\overline{\text{PMI}}_3 = 0.05$ bits (bootstrap 95% CIs exclude zero for $k = 1, 2$ but not for $k = 3$), indicating a monotonically decaying but, at distance two, still clearly non-trivial dependency structure between non-adjacent sign positions. This is consistent with the case-internal slot-grammar account (H2) under which a commodity sign at one slot constrains, but does not rigidly determine, the office or personnel sign two slots downstream, and provides additional n-gram-level support, independent of the contiguous-bigram evidence, for a combinatorially structured rather than purely linear generative process.

4.9 Case-Internal Versus Cross-Case Transition Asymmetry

Proto-Elamite tablets are physically organized into discrete sealed compartments ("cases"), and a natural question — without a direct analogue in the Linear A textual record, which lacks comparable physical compartmentalization — is whether the sequential statistics estimated in Sections 4.1–4.7 differ depending on whether a sign transition occurs within a single case or crosses a case boundary.

Proposition 4.4: Case Boundaries Are Statistically Salient Transition Points

Partitioning all observed bigram transitions into case-internal ($n = 14{,}772$) and case-crossing ($n = 3{,}118$) subsets and re-estimating the first-order transition entropy separately for each yields $\hat{H}_1^{\text{internal}} = 5.41$ bits and $\hat{H}_1^{\text{crossing}} = 6.87$ bits, a difference of 1.46 bits (95% bootstrap CI $[1.21, 1.69]$, non-overlapping). Case-crossing transitions are substantially less predictable than case-internal ones, exactly as expected if cases function as the primary semantic/grammatical unit of the notation: knowledge of a sign strongly constrains the identity of the next sign *within the same case* but only weakly constrains the first sign of the *following* case, which typically opens a fresh, largely independent administrative entry.

Corollary 4.1: Case Boundaries as an Independent Segmentation Signal

Because case boundaries are physically marked on the tablet (by sealing and ruled lines) independently of any distributional inference, Proposition 4.4 provides an external validation of the boundary-inference statistic of Definition 3.9: the case boundary, where physically attested, coincides with a local maximum of the boundary-strength statistic B_i in 84% of instances ($n = 1{,}847$ physically marked case boundaries with adequate surrounding transcription), substantially above the 30% co-occurrence rate expected if B_i were uninformative about physical case structure.

© Silvano Salvador

Chapter 5

The Reference Corpus and Formal Sign-System Classification

5.1 Constructing a Typologically Diverse Reference Set

Definition 5.1: The Reference Corpus $\mathcal{R}$

The reference corpus $\mathcal{R}$ is the union of six typologically distinct sign systems, digitized and tokenized using a uniform protocol for comparison against the proto-Elamite non-numerical sub-corpus $\mathcal{S}$:

- $\mathcal{R}_1$: Proto-cuneiform (Uruk IV–III), $\approx$ 5,000 tablets, $\approx$ 137,000 sign tokens, a known logographic-numerical administrative notation ancestral to Sumerian cuneiform;
- $\mathcal{R}_2$: Linear Elamite, $\approx$ 40 inscriptions, $\approx$ 1,600 sign tokens, a partially-read syllabic-logographic successor script (see Section 8.2);
- $\mathcal{R}_3$: Linear B, $\approx$ 6,000 tablets, $\approx$ 60,000 sign tokens, a fully deciphered syllabic script encoding Mycenaean Greek, included as a positive control for genuine linguistic encoding;
- $\mathcal{R}_4$: Indus Valley sign sequences, $\approx$ 4,200 inscriptions, $\approx$ 13,000 sign tokens, a formally undeciphered system of disputed linguistic status, included as a second undeciphered comparandum;
- $\mathcal{R}_5$: Synthetic pure-tallying control, 10,000 generated tokens drawn from a single repeated symbol with positional noise, simulating a degenerate one-symbol counting system;
- $\mathcal{R}_6$: Synthetic constrained-semasiography control, 10,000 generated tokens drawn from a hand-built generative grammar of 120 symbols organized into commodity/quantity/institution slot classes, simulating a non-linguistic but combinatorially structured notation.

All six reference systems are tokenized at the sign level and partitioned into case- or line-level sequence units directly analogous to the proto-Elamite administrative case structure of Definition 1.2.

Note 5.1: Comparability Caveats

Direct comparison across $\mathcal{R}_1$–$\mathcal{R}_6$ is complicated by substantial differences in corpus size, ranging from $N_{\mathcal{R}_2} \approx 1{,}600$ to $N_{\mathcal{R}_1} \approx 137{,}000$ tokens. All entropy and Zipf es-

timates reported in this chapter are accompanied by bootstrap confidence intervals computed at matched sample sizes (subsampling larger corpora down to $N_S \approx 10{,}000$ tokens where direct comparison to proto-Elamite is intended), in order to avoid spurious differences attributable purely to estimator variance at small N.

5.2 Formal Criteria for Logographic, Syllabic, and Semasiographic Systems

Definition 5.2: The Three Discriminating Statistics

Three corpus-level statistics are formalized as candidate discriminators between logographic, syllabic, and semasiographic sign systems:

1. **Sign-inventory size** $|\Sigma|$: syllabic systems typically exhibit $|\Sigma| \in [40, 100]$ (a phonologically plausible syllable inventory); logographic systems exhibit $|\Sigma| \in [500, 3000]$ (one sign per morpheme or concept class); semasiographic systems are typically intermediate, $|\Sigma| \in [100, 700]$, reflecting a constrained but non-exhaustive concept inventory.
2. **Entropy plateau height** H_∞: the asymptotic value of the conditional entropy H_k as $k \to \infty$ (Definition 3.4), estimated in practice via extrapolation from H_1–H_3. Natural languages, whether logographic or syllabic in script, exhibit $H_\infty/H_0 \in [0.25, 0.45]$ (substantial redundancy reduction at higher context orders); semasiographic and tallying systems exhibit higher ratios, $H_\infty/H_0 \in [0.55, 0.85]$, reflecting weaker long-range sequential dependency.
3. **Positional regularity** ρ_{pos}: the normalized mutual information between sign identity and within-case position, $\rho_{pos} = I(S; \text{Pos})/H(S)$. High ρ_{pos} indicates a slot-grammar structure (characteristic of semasiographic systems with fixed commodity/quantity/agent ordering); low ρ_{pos} is characteristic of free word-order or syllabically-encoded natural language.

Finding 5.1: The Three Statistics Computed Across $\mathcal{R}$

System	$\lvert\Sigma\rvert$	H_∞/H_0	ρ_{pos}
Proto-cuneiform ($\mathcal{R}_1$)	1,500	0.61	0.34
Linear Elamite ($\mathcal{R}_2$)	91	0.31	0.09
Linear B ($\mathcal{R}_3$)	87	0.27	0.07
Indus Valley ($\mathcal{R}_4$)	417	0.58	0.29
Pure tallying ($\mathcal{R}_5$)	3	0.94	0.02
Constrained semasiography ($\mathcal{R}_6$)	120	0.68	0.41
Proto-Elamite ($\mathcal{S}$)	**637**	**0.66**	**0.38**

The proto-Elamite profile ($|\Sigma| = 637$, $H_\infty/H_0 = 0.66$, $\rho_{pos} = 0.38$) falls closest, across all three statistics jointly, to the constrained-semasiography control $\mathcal{R}_6$ and to proto-cuneiform $\mathcal{R}_1$, and is well separated from both syllabic controls ($\mathcal{R}_2$, $\mathcal{R}_3$, with $H_\infty/H_0 < 0.35$) and from the degenerate tallying control $\mathcal{R}_5$. Euclidean distance in standardized three-statistic space confirms this ordering: $d(\mathcal{S}, \mathcal{R}_6) = 0.41$, $d(\mathcal{S}, \mathcal{R}_1) = 0.53$, versus $d(\mathcal{S}, \mathcal{R}_3) = 1.87$.

© Silvano Salvador

5.3 Where Proto-Elamite Falls in the Classification Space

Proposition 5.1: Confidence Region in Classification Space

Bootstrap resampling of the proto-Elamite corpus ($B = 2{,}000$ resamples at the case level) propagates sampling uncertainty through the three discriminating statistics, yielding a joint 90% confidence ellipsoid in $(|\Sigma|, H_\infty/H_0, \rho_{pos})$-space. The ellipsoid is centered on the point estimate of Finding 5.1 and has semi-axis lengths of ± 38 (sign inventory), ± 0.07 (H_∞/H_0), and ± 0.06 (ρ_{pos}). The constrained-semasiography control $\mathcal{R}_6$ and proto-cuneiform $\mathcal{R}_1$ both fall within or adjacent to this ellipsoid; both syllabic controls fall well outside it at all conventional confidence levels.

Corollary 5.1: A Provisional Classification

On the strength of Finding 5.1 and Proposition 5.1 alone, the proto-Elamite non-numerical sign system is provisionally classified as structurally closer to a constrained semasiography or to a proto-cuneiform-like proto-logographic administrative notation than to a syllabically- or logographically-encoded full natural language. This classification is provisional in two respects: first, it is based on corpus-level statistics that cannot rule out a register-restricted natural language (Section 3.3's restricted-register anomaly remains formally compatible with H1 under sufficiently narrow registers); second, the confidence ellipsoid of Proposition 5.1 does not exclude proto-cuneiform-like structure, and proto-cuneiform itself, while well understood administratively, encoded only a partial and evolving relationship to spoken Sumerian in its earliest (Uruk IV) phase. The formal hypothesis evaluation of Chapter 6 is required to adjudicate among the remaining possibilities.

Worked Example 5.1: Locating Indus Valley Signs for Context

The Indus Valley system ($\mathcal{R}_4$), itself formally undeciphered, falls at $d(\mathcal{S}, \mathcal{R}_4) = 0.49$ from the proto-Elamite point estimate — comparably close to $\mathcal{R}_1$ and $\mathcal{R}_6$. This proximity is noted but not interpreted as evidence of any historical relationship between the two systems, which are separated by over a thousand kilometers and several centuries; rather, it illustrates that the three-statistic classification space groups together typologically similar *undeciphered administrative notations* regardless of geographic or chronological relationship, a pattern consistent with convergent development of constrained semasiographic or proto-logographic systems under similar administrative pressures.

5.4 Extending the Classification Space: Additional Discriminating Statistics

The three-statistic classification space of Definition 5.2 — $|\Sigma|$, H_∞/H_0, and ρ_{pos} — provides the primary typological placement results of Section 5.3. This section extends the classification with two additional statistics that provide complementary discrimination, particularly between the T3 (constrained semasiography) and T4 (proto-logographic) types of Definition G.1.

© Silvano Salvador

Definition 5.3: The Lexical Diversity Index D_{lex}

The **lexical diversity index** is defined as

$$D_{lex} = \frac{|\Sigma^*|}{N^{0.67}},$$

the ratio of observed vocabulary size to the two-thirds power of corpus size. This normalizes vocabulary growth by Heaps' law expectation ($|\Sigma| \propto N^{0.67}$ for many natural-language corpora), yielding a measure of *relative* vocabulary richness. A higher D_{lex} indicates greater vocabulary richness relative to corpus size, expected for sign systems encoding a wider range of distinct concepts or linguistic units. For proto-Elamite: $D_{lex} = 637/(10{,}142)^{0.67} = 637/461 = 1.38$; for Linear B ($\mathcal{R}_3$, $N = 60{,}000$, $|\Sigma| = 87$): $D_{lex} = 87/(60{,}000)^{0.67} = 87/1{,}649 = 0.053$.

Proposition 5.2: D_{lex} Discriminates Semasiographic from Syllabic Systems

Across the six reference systems, the lexical diversity index ranges from 0.053 (Linear B, $\mathcal{R}_3$) to 2.14 (synthetic semasiography, $\mathcal{R}_6$), with proto-Elamite at 1.38 and proto-cuneiform at 1.61 — both substantially above the syllabic controls ($D_{lex}^{LA} = 0.44$ for Linear Elamite, $\mathcal{R}_2$) and below the synthetic semasiographic control. The separation between syllabic systems ($D_{lex} < 0.50$) and semasiographic/proto-logographic systems ($D_{lex} > 1.0$) is cleaner than the separation achieved by either H_∞/H_0 or ρ_{pos} individually, making D_{lex} a valuable additional discriminating statistic for the sign-system classification problem despite its simplicity.

Definition 5.4: The Case-Length Entropy H_{CL}

The **case-length entropy** is the Shannon entropy of the distribution of administrative case lengths (number of sign tokens per case):

$$H_{CL} = -\sum_{\ell=1}^{\ell_{max}} p(\ell) \log_2 p(\ell),$$

where $p(\ell)$ is the empirical proportion of cases with length ℓ tokens. A low H_{CL} indicates highly regular case lengths (consistent with a rigid slot-grammar template); a high H_{CL} indicates highly variable case lengths (consistent with free-length natural-language clauses). For proto-Elamite: $H_{CL} = 2.14$ bits over a case-length range of 1–11 tokens; for Linear B: $H_{CL} = 2.88$ bits; for the synthetic semasiography control $\mathcal{R}_6$: $H_{CL} = 1.71$ bits.

Proposition 5.3: Five-Statistic Classification Confirms the H2 Placement

Adding D_{lex} and H_{CL} to the three-statistic classification of Proposition 5.1, the five-statistic Euclidean distance in standardized space from proto-Elamite to each reference system: $d(\mathcal{S}, \mathcal{R}_6) = 0.38$, $d(\mathcal{S}, \mathcal{R}_1) = 0.51$, $d(\mathcal{S}, \mathcal{R}_4) = 0.44$, $d(\mathcal{S}, \mathcal{R}_3) = 2.04$, $d(\mathcal{S}, \mathcal{R}_2) = 1.98$, $d(\mathcal{S}, \mathcal{R}_5) = 3.11$. The ordering is identical to the three-statistic result (Finding 5.1), confirming that the classification is robust to the choice of statistic set within the range examined here. The constrained-semasiography control $\mathcal{R}_6$ remains the closest reference system at the five-statistic level, strengthening the H2-consistent

© Silvano Salvador

typological placement of Corollary 5.1.

5.5 The Sign-System Classification Space as a Research Tool

Note 5.2: Toward a Universal Undeciphered-Script Classification Framework

The three-plus-two-statistic classification space developed in this chapter and Appendix G constitutes a practical tool for the systematic typological placement of undeciphered scripts. Its key properties are: (i) it uses only *corpus-internal* statistics, requiring no external knowledge of the script's referential content; (ii) it produces a continuous vector representation (the point in the classification space) that allows gradations between types, not just binary type-assignments; and (iii) it is anchored by known reference systems at both ends of the typological range (Linear B as a syllabic extreme, the synthetic tallying control $\mathcal{R}_5$ as the degenerate extreme), so the classification is calibrated against systems of known status. Applied to the three scripts of this monograph series (Voynich, Linear A, proto-Elamite), the framework produces three distinct placements in the classification space — evidence that it responds to genuine structural differences among the three corpora rather than generating a uniform artifact.

Worked Example 5.2: Where the Voynich Manuscript Falls in This Framework

For comparison with the companion Voynich monograph, the Voynich manuscript's non-Currier A word-sequence corpus (the portion most analogous to the proto-Elamite non-numerical sub-corpus) has the following five-statistic profile, estimated from the EVA transcription: $|\Sigma| \approx 280$ word types (using words rather than letter sequences as the unit of analysis), $H_\infty/H_0 \approx 0.60$, $\rho_{pos} \approx 0.25$, $D_{lex} \approx 0.74$, $H_{CL} \approx 2.40$ bits. The Euclidean distance from Voynich to proto-Elamite in the five-statistic space is $d(\text{Voynich}, \mathcal{S}) = 0.92$ — substantially further from proto-Elamite than any of the six reference systems listed in Proposition 5.3, and falling in a region of the space consistent with a Type T3–T4 placement (like proto-Elamite) but with a somewhat lower positional regularity and a smaller effective sign inventory. This cross-script comparison is offered for context, not as a claim of typological kinship between the two systems.

5.6 Robustness of the Placement to Reference-Set Composition

A natural concern about any typological placement exercise is that the result may be an artifact of the particular reference systems chosen to anchor the classification space (Definition 5.1) rather than a robust property of the proto-Elamite corpus itself. This section addresses that concern directly via a leave-one-out sensitivity analysis.

Proposition 5.4: Leave-One-Reference-System-Out Stability

Recomputing proto-Elamite's five-statistic placement (Proposition 5.3) after removing each reference system from $\mathcal{R}$ in turn, refitting the classification-space axes, and reprojecting proto-Elamite into the reduced space yields a maximum placement shift of 0.09 (in normalized Euclidean distance) across all seven leave-one-out refits, with

© Silvano Salvador

the largest shift occurring when Linear B — the system contributing the most extreme syllabic anchor point — is removed. All seven leave-one-out placements remain within the Type T3 region of the classification space (Definition 5.2's discriminating statistics, applied to the boundary contours of Corollary 5.1), confirming that the H2-consistent placement of Corollary 5.1 is not an artifact of any single reference system's inclusion.

Note 5.3: Why Linear B's Removal Has the Largest (but Still Small) Effect

Linear B's outsized leverage on the classification-space axes is expected on purely geometric grounds: as the system closest to the syllabic extreme of the typological range, it does the most work in calibrating the syllabic pole against which proto-Elamite's distance is measured (Note 5.2, property iii). That its removal shifts the placement by less than a tenth of the space's characteristic scale, rather than substantially relocating proto-Elamite, indicates that the classification is being driven primarily by proto-Elamite's own corpus-internal statistics and only secondarily by the precise calibration of the reference anchors.

5.7 Administrative and Lexical Sign Sub-Classes: A Finer-Grained Partition

The five-statistic classification of Sections 5.3–5.4 treats the non-numerical sign inventory $\mathcal{S}$ as a single undifferentiated population. Cuneiformist practice (Englund, 2004; Dahl, 2005), however, distinguishes within $\mathcal{S}$ between *administrative* signs (measure-words, classifiers, and abbreviation-like notations recurring across many commodity types) and *lexical* signs (commodity, office, toponym, and personal-name signs with more restricted, content-specific distributions). This section tests whether the two sub-classes occupy distinguishable regions of classification space.

Definition 5.5: The Administrative/Lexical Partition Criterion

A sign is provisionally assigned to the administrative sub-class $\mathcal{S}_{\text{adm}}$ if it co-occurs with at least 15 distinct line-initial commodity signs at a frequency exceeding chance (binomial test, $\alpha = 0.01$, Bonferroni-corrected across the sign inventory), and to the lexical sub-class $\mathcal{S}_{\text{lex}}$ otherwise. The procedure classifies 164 of the $\approx$ 1,000 attested non-numerical signs as administrative and the remainder as lexical (with 612 signs occurring too infrequently, $n < 5$ tokens, to classify reliably and left unassigned).

Finding 5.2: A Sharp Classification-Space Split Between Sub-Classes

Recomputing the five-statistic profile separately for $\mathcal{S}_{\text{adm}}$ and $\mathcal{S}_{\text{lex}}$ (restricting to the $n = 5{,}716$ and $n = 11{,}284$ tokens respectively with adequate per-class token counts for stable estimation) places the administrative sub-class substantially closer to the synthetic tallying control $\mathcal{R}_5$ ($d = 0.31$) than the lexical sub-class ($d = 0.88$), while the lexical sub-class alone lands almost exactly at the pooled corpus placement reported in Proposition 5.3 ($d = 0.06$ from the pooled estimate). This indicates that the combinatorial richness driving the overall H2-consistent classification is disproportionately carried by the lexical sign sub-class, while the administrative sub-class behaves closer to a constrained, near-formulaic notation — a finding that sharpens, rather than con-

© Silvano Salvador

tradicts, the slot-grammar account of Chapter 4, under which administrative signs are expected to fill a small number of recurring grammatical roles while lexical signs carry the bulk of the combinatorial content.

© Silvano Salvador

Chapter 6

Evaluating the Principal Hypotheses

Note 6.1: Evaluation Protocol

Each of the four hypotheses H1–H4 introduced in Definition 1.3 is evaluated here against a common battery of six diagnostics established in Chapters 3–5: (a) the entropy plateau ratio H_∞/H_0; (b) the Zipf exponent α_Z; (c) the positional regularity ρ_{pos}; (d) the effective model-order ceiling k^*; (e) the cross-entropy distance to proto-cuneiform; and (f) the cross-site transfer-entropy asymmetry. For each hypothesis, the chapter states the diagnostic predictions the hypothesis makes *a priori*, then scores the degree of match against the values established empirically in Chapters 3–5. Formal Bayes factors and MDL scores synthesizing this evaluation are deferred to Chapter 7.

6.1 H1: A Natural-Language Encoding

Definition 6.1: H1 — Formal Statement

H1 holds that the non-numerical proto-Elamite signs encode, in whole or in significant part, a spoken natural language — most plausibly an archaic predecessor or relative of Elamite, the only well-attested indigenous language of the region in later periods. Under H1, sign sequences within an administrative case correspond, at least loosely, to syntactic or morphological structure in the underlying language, and the sign-frequency distribution should approximate the statistical signature of natural-language text rather than that of a closed administrative notation.

Proposition 6.1: A Priori Diagnostic Predictions Under H1

If H1 holds, the diagnostic battery should exhibit: (a) H_∞/H_0 substantially below 0.5, typically in $[0.25, 0.45]$ as observed for Linear B and Linear Elamite; (b) a Zipf exponent α_Z close to the canonical natural-language value of approximately 1.0–1.1; (c) low positional regularity, $\rho_{pos} \lesssim 0.15$, reflecting syntactic rather than slot-grammar ordering; (d) an effective model-order ceiling $k^* \geq 2$, since natural-language sequences typically exhibit detectable trigram-level structure even at modest corpus sizes; and (e) low cross-entropy to other natural-language-encoding comparanda ($\mathcal{R}_2$, $\mathcal{R}_3$) relative to non-linguistic comparanda.

Finding 6.1: H1 Is Weakly Supported by the Diagnostic Battery

The observed values are: $H_\infty/H_0 = 0.66$ (Finding 5.1, contrary to the H1 prediction of $[0.25, 0.45]$); $\alpha_Z = 0.78$ (Finding 3.2, contrary to the natural-language prediction of $\approx$ 1.0–1.1); $\rho_{pos} = 0.38$ (contrary to the H1 prediction of $\lesssim 0.15$); $k^* = 1$ (Proposition 4.2, contrary to the H1 prediction of $k^* \geq 2$, though this may partly reflect corpus-size limitations rather than a genuine absence of higher-order structure). Of the five diagnostic predictions, none is satisfied at the point-estimate level, though the k^* result carries the important caveat that the Voynich comparandum (also $k^* = 1$ at comparable size) is independently suspected of encoding meaningful structure, weakening k^* as a standalone discriminator. On balance, H1 in its unrestricted form is weakly supported by the corpus-internal evidence.

Note 6.2: The Restricted-Register Escape Clause

A narrower version of H1 — under which the proto-Elamite corpus encodes natural language but in a highly restricted administrative register, analogous to a shopping list or invoice rather than free prose — remains formally compatible with several of the anomalous diagnostic values, since restricted registers are independently known to exhibit reduced entropy and inflated positional regularity relative to free text (the restricted-register anomaly of Section 3.3). This version of H1 cannot be excluded by the present diagnostic battery and is carried forward into the Bayesian synthesis of Chapter 7 as a distinct sub-hypothesis, H1′.

6.2 H2: Constrained Administrative Semasiography

Definition 6.2: H2 — Formal Statement

H2 holds that the non-numerical signs constitute a constrained semasiography: a combinatorially structured notation in which signs denote commodities, institutions, personnel categories, or transaction types directly, without mediating through the phonology or full syntax of any spoken language. Under H2, sign co-occurrence is governed by a slot-grammar (commodity + qualifier + institution, or similar fixed templates) rather than by linguistic syntax, and sign-frequency structure should resemble that of the constrained-semasiography control $\mathcal{R}_6$ and, to a lesser extent, proto-cuneiform $\mathcal{R}_1$.

Proposition 6.2: A Priori Diagnostic Predictions Under H2

If H2 holds, the diagnostic battery should exhibit: (a) H_∞/H_0 elevated relative to natural language, in $[0.55, 0.85]$; (b) a Zipf exponent below the natural-language canonical range, $\alpha_Z \in [0.6, 0.9]$, reflecting a smaller effective "vocabulary" of frequently-reused administrative categories; (c) elevated positional regularity, $\rho_{pos} \gtrsim 0.30$, reflecting slot-grammar structure; (d) a low effective model-order ceiling, $k^* \in \{0, 1\}$, since slot grammars are typically well captured by low-order Markov structure; and (e) low cross-entropy to proto-cuneiform and to the synthetic semasiography control $\mathcal{R}_6$.

© Silvano Salvador

Finding 6.2: H2 Is Strongly Supported by the Diagnostic Battery

All five diagnostic predictions are satisfied at the point-estimate level: $H_\infty/H_0 = 0.66 \in [0.55, 0.85]$; $\alpha_Z = 0.78 \in [0.6, 0.9]$; $\rho_{pos} = 0.38 \gtrsim 0.30$; $k^* = 1 \in \{0, 1\}$; and the cross-entropy distance to $\mathcal{R}_6$ ($d = 0.41$, Finding 5.1) is the smallest of any reference system. H2 is the only one of the four hypotheses under formal evaluation in this chapter whose full diagnostic prediction set is satisfied without qualification by the empirical results of Chapters 3–5.

6.3 H3: Personal-Name and Institutional Identifiers

Definition 6.3: H3 — Formal Statement

H3 holds that a substantial fraction of the non-numerical signs — though not necessarily all of them — function as identifying labels (personal names, institutional affiliations, or place names) attached to an otherwise numerical accounting framework, analogous to the function of personal-name determinatives and place-name signs in early proto-cuneiform and in the Linear B "sa" and toponym series. Under H3, a minority subset of signs should exhibit high positional fixity (appearing overwhelmingly in a single case-position) and low co-occurrence entropy with neighboring numerical notations, while the majority of non-numerical signs need not exhibit this pattern.

Proposition 6.3: Sign-Recurrence and Positional-Clustering Diagnostics

H3 is tested via two additional sign-level (rather than corpus-level) statistics not used in the H1/H2 evaluation: the *positional fixity index* $\phi_s = \max_k \hat{p}(\text{Pos} = k \mid s)$ for each sign s, and the *numerical co-occurrence rate* ν_s, the proportion of occurrences of s immediately adjacent to a numerical notation token. H3 predicts a bimodal distribution of ϕ_s across the sign inventory: a subset of signs (candidate identifiers) with $\phi_s > 0.7$ and $\nu_s > 0.5$, and a remaining majority with ϕ_s and ν_s closer to their corpus-wide means.

Finding 6.3: H3 Is Supported for a Minority Sign Subset

The empirical distribution of ϕ_s across the 637-sign inventory is bimodal (Hartigan's dip test, $D = 0.041$, $p = 0.003$, rejecting unimodality), with a subset of $n = 73$ signs (11.5% of the inventory, accounting for 8.2% of non-numerical tokens) exhibiting $\phi_s > 0.7$ and $\nu_s > 0.5$, consistent with an identifier function. The remaining 564 signs exhibit ϕ_s and ν_s distributions statistically indistinguishable from the corpus-wide background ($\chi^2 = 1.84$, $p = 0.40$). This finding supports H3 as a *partial* account: a minority sign subset plausibly functions as personal-name or institutional identifiers, but the majority of the non-numerical sign inventory is not well explained by H3 alone, and the diagnostic profile of this majority subset continues to match the H2 predictions of Proposition 6.2 closely when the candidate-identifier subset is excluded.

Worked Example 6.1: A Candidate Identifier Sign

Sign PE-211 occurs 34 times in the digitized corpus, of which 31 occurrences ($\phi_{\text{PE-211}} = 0.91$) fall in case-final position, and 29 occurrences ($\nu_{\text{PE-211}} = 0.85$) are immediately preceded by a livestock-system numerical notation. This profile — high positional fixity,

© Silvano Salvador

strong numerical co-occurrence, case-final placement — is consistent with a function as either a recipient/institution label or an animal-class qualifier; the corpus-internal statistics alone cannot distinguish between these two readings, illustrating the general limit that sign-level statistical diagnostics can identify functional *categories* of sign behavior without resolving specific semantic content.

6.4 H4: Localized, Non-Standardized Notation

Definition 6.4: H4 — Formal Statement

H4 holds that the apparent combinatorial structure identified in Chapters 3–5 reflects site-specific or scribe-specific convention rather than a single shared system standardized across the proto-Elamite cultural sphere. Under H4, cross-site comparisons should reveal substantially divergent sign-frequency distributions and weak or absent transfer entropy between sites, since each site's "system" would be, in the relevant sense, a locally idiosyncratic notation rather than a shared code.

Proposition 6.4: Diagnostic Predictions Under H4

If H4 holds, the diagnostic battery should exhibit: (a) high cross-site KL divergence in sign-frequency distributions, comparable in magnitude to the proto-Elamite/proto-cuneiform divergence already established (Finding 4.1, D_{KL} sum $= 4.29$ bits); and (b) transfer entropy asymmetries $\Delta T_{Site_1 \to Site_2}$ statistically indistinguishable from zero, since no consistent directional influence should exist between genuinely independent local notations.

Finding 6.4: H4 Is Not Supported by the Cross-Site Evidence

Contrary to the H4 prediction, the Susa/Anshan cross-site KL divergence sum is $D_{KL}(\text{Susa}\|\text{Anshan}) + D_{KL}(\text{Anshan}\|\text{Susa}) = 0.34$ bits, an order of magnitude smaller than the proto-Elamite/proto-cuneiform divergence of 4.29 bits, indicating that Susa and Anshan share a substantially common sign-frequency structure. Furthermore, the transfer-entropy asymmetry $\Delta T_{Susa \to Anshan} = 0.043$ bits established in Finding 4.2 is significantly positive ($p = 0.01$), directly contradicting the H4 prediction of statistically indistinguishable directional influence. H4 is accordingly rejected as a general account of the corpus, though it is retained in a weakened form restricted to the documented PE III fragmentation (Finding 2.1, Note 4.1), where cross-phase transfer entropy does become statistically undetectable, consistent with site- or workshop-level divergence emerging specifically in the latest phase.

6.5 A Joint Diagnostic Score and Its Sensitivity to Weighting

Corollary 6.1's ranking is reached qualitatively, by inspecting each hypothesis's diagnostic battery in turn. This section formalizes the comparison as a single weighted composite score and tests whether the resulting ranking is sensitive to the (necessarily somewhat arbitrary) choice of weights.

© Silvano Salvador

Definition 6.5: The Composite Diagnostic Score

For hypothesis H_i and diagnostic statistic $j \in \{1, \ldots, 5\}$ (the five statistics of Definition 5.2 and Definition 5.3/5.4), let z_{ij} be the standardized distance between the observed corpus value and H_i's predicted value (in units of the predicted value's own uncertainty, where available, or of the cross-reference-system spread otherwise). The composite score is

$$\Sigma_i = \sum_{j=1}^{5} w_j \, z_{ij}, \qquad \sum_j w_j = 1, \; w_j \geq 0,$$

with lower Σ_i indicating better diagnostic agreement.

Proposition 6.5: The H2 Ranking Is Robust to Weighting Choice

Under equal weighting ($w_j = 0.2$ for all j), the composite scores are $\Sigma_{H1} = 2.81$, $\Sigma_{H2} = 0.64$, $\Sigma_{H3} = 1.92$ (restricted to the identifier-eligible sign subset), $\Sigma_{H4} = 2.15$, placing H2 lowest (best) by a wide margin. Re-running the comparison under nine alternative weighting schemes — including weight concentrated entirely on any single diagnostic statistic, and a leave-one-statistic-out scheme analogous to Proposition 5.4 — leaves H2 ranked lowest in eight of nine schemes; the sole exception concentrates all weight on the model-order ceiling k^* alone, a statistic explicitly flagged in Finding 6.1 as a weak discriminator on its own. The composite ranking is accordingly judged robust to reasonable variation in diagnostic weighting, rather than an artifact of any single chosen scheme.

6.6 Out-of-Sample Validation: Holding Out Tepe Yahya

A further robustness check, independent of the weighting-sensitivity analysis above, asks whether the qualitative ranking of Corollary 6.1 is an artifact of the Susa-dominated composition of the corpus (Susa contributes $\approx 78\%$ of all tokens) rather than a genuinely corpus-wide pattern.

Worked Example 6.2: Re-Deriving the Ranking on Tepe Yahya Alone

Restricting the full diagnostic battery to the $n \approx 1{,}340$ tokens recovered from Tepe Yahya (the largest non-Susa findspot, and the one most plausibly reflecting locally autonomous administrative practice given its peripheral location relative to the Susa–Anshan core) and recomputing the composite score of Definition 6.5 under equal weighting yields $\Sigma_{H1} = 2.64$, $\Sigma_{H2} = 0.81$, $\Sigma_{H3} = 2.05$, $\Sigma_{H4} = 1.97$: H2 remains the lowest-scoring (best-supported) hypothesis, though its margin over H4 narrows relative to the full-corpus result, consistent with Tepe Yahya's greater geographic and administrative distance from the Susa core plausibly admitting a somewhat larger admixture of locally idiosyncratic notation.

Note 6.3: Interpreting the Narrowed Margin

The narrowed H2/H4 margin at Tepe Yahya should not be over-interpreted as evidence *for* H4: the formal rejection of H4 in Finding 6.4 rests on the transfer-entropy asymme-

© Silvano Salvador

try between Susa and Anshan, a comparison not re-derivable on the single-site Tepe Yahya sub-corpus for want of a second site against which to measure transfer entropy. The Tepe Yahya result is better read as showing that the H2-consistent slot-grammar structure, while dominant, is not uniquely a property of the largest (Susa) sub-corpus, and that some genuine site-level heterogeneity — already documented in Finding 2.1's phase-fragmentation result — persists at the periphery of the proto-Elamite administrative network.

Corollary 6.1: A Preliminary Ranking Ahead of Formal Bayesian Synthesis

The qualitative evaluation of this chapter yields a preliminary, informal ranking: H2 (constrained administrative semasiography) is most strongly supported by the diagnostic battery considered as a whole, and this support is robust both to the choice of diagnostic weighting (Proposition 6.5) and to restriction to a single peripheral site (Worked Example 6.2); H3 (identifier hypothesis) is supported as a partial account applicable to a minority sign subset, and is not in direct competition with H2 for the majority of the corpus; H1 (natural-language encoding) is weakly supported in its unrestricted form but cannot be excluded in its restricted-register variant H1′; and H4 (localized notation) is rejected as a general account, surviving only as a phase-restricted qualification on H2. This informal ranking is formalized via Bayes factors and MDL scoring in Chapter 7.

© Silvano Salvador

Chapter 7

Bayesian Model Comparison and the MDL Synthesis

7.1 Prior Elicitation for the Five Hypotheses

Note 7.1: From Four Hypotheses to Five

Following Note 6.2, the restricted-register variant of H1 is carried forward as an independent hypothesis H1′ in this chapter's formal synthesis, alongside H1, H2, H3, and H4, for a total of five candidate generative accounts. H3, as established in Finding 6.3, is treated not as a competitor to H2 for the bulk of the corpus but as a compatible sub-structure; the Bayesian synthesis below accordingly evaluates H3 as a hypothesis about the minority identifier subset specifically, conditioned on H2 holding for the corpus majority.

Definition 7.1: Elicited Priors

Priors $\pi(H_i)$ are elicited via a structured survey of the published epigraphic and computational-epigraphy literature on proto-Elamite (weighting recent quantitative treatments above purely qualitative ones), combined with reference-class priors transferred from the analogous Voynich hypothesis space (Section 7.1 of the companion Voynich monograph), adjusted for the cuneiform-adjacent administrative context that distinguishes the proto-Elamite case. The elicited priors are:

$$\pi(H1) = 0.12, \quad \pi(H1') = 0.18, \quad \pi(H2) = 0.40, \quad \pi(H3) = 0.10, \quad \pi(H4) = 0.20.$$

The comparatively high prior on H2 reflects the strong typological and contextual similarity to proto-cuneiform established in Chapter 5; the non-trivial prior retained on H4 reflects the genuine cross-site divergence documented in PE III (Note 4.1) even though full-corpus H4 is rejected by Finding 6.4.

Note 7.2: Prior Sensitivity Analysis

A sensitivity analysis recomputes the final posterior ranking under three alternative prior specifications: a uniform prior ($\pi(H_i) = 0.2$ for all i), an "optimistic decipherment" prior favoring H1/H1′ ($\pi(H1) = \pi(H1') = 0.30$, remaining mass split uniformly), and a "skeptical" prior favoring H4 ($\pi(H4) = 0.45$, remaining mass split

proportionally to the elicited prior). The posterior ranking of Section 7.4 is invariant to all three alternative specifications, with H2 retaining the highest posterior probability in every case; only the *magnitude* of the posterior gap between H2 and its nearest competitor varies materially across prior specifications, narrowing to a factor of 2.1 under the optimistic-decipherment prior and widening to a factor of 11.4 under the skeptical prior.

7.2 Bayes Factor Computation Across the Diagnostic Battery

Definition 7.2: The Composite Likelihood

For each hypothesis H_i, a composite log-likelihood is constructed as a weighted sum over the six diagnostics of Note 6.1, with weights reflecting each diagnostic's estimated discriminating power (inverse-variance weighting based on the bootstrap confidence intervals of Proposition 5.1 and the corresponding intervals computed for the Zipf, cross-entropy, and transfer-entropy statistics):

$$\log \hat{\mathcal{L}}(H_i) = \sum_{j=1}^{6} w_j \log p_j(\text{data} \mid H_i),$$

where each $p_j(\cdot \mid H_i)$ is modeled as a Gaussian density over the diagnostic's predicted range under H_i (Propositions 6.1–6.4), centered at the midpoint of the predicted interval with standard deviation set to one-quarter of the interval width, evaluated at the observed point estimate.

Finding 7.1: Pairwise Bayes Factors

	H1	H1′	H2	H3	H4
H1	—	0.21	0.04	0.31	0.58
H1′	4.76	—	0.19	1.48	2.76
H2	24.1	5.26	—	7.79	14.5
H3	3.23	0.68	0.13	—	1.86
H4	1.72	0.36	0.069	0.54	—

Cell (i, j) reports the Bayes factor $K_{ij} = \hat{\mathcal{L}}(H_i)/\hat{\mathcal{L}}(H_j)$ (row hypothesis over column hypothesis). H2 dominates all four competitors, with Bayes factors ranging from $K_{H2,H1} = 24.1$ ("decisive" evidence on the conventional Jeffreys/Kass–Raftery scale) to $K_{H2,H1'} = 5.26$ ("strong" evidence). H1′ in turn dominates unrestricted H1 ($K_{H1',H1} = 4.76$, "substantial" evidence), confirming that the restricted-register reformulation materially improves H1's fit to the diagnostic battery without closing the gap to H2.

© Silvano Salvador

7.3 The MDL Differential and Its Interpretation

Definition 7.3: MDL Differential

For each hypothesis H_i, the Minimum Description Length differential $\Delta\text{MDL}(H_i)$ relative to a naive baseline (fixed-length binary encoding of all 637 signs, requiring $\lceil \log_2 637 \rceil = 10$ bits per token) is

$$\Delta\text{MDL}(H_i) = N_{\mathcal{S}} \times 10 - [L(\text{model}_i) + L(\text{data} \mid \text{model}_i)],$$

where $L(\text{model}_i)$ is the description length of the hypothesis's generative model (parameter count weighted by per-parameter encoding cost) and $L(\text{data} \mid \text{model}_i)$ is the negative log-likelihood of the observed corpus under that model, both in bits. A larger ΔMDL indicates a more compressively efficient account of the data relative to the naive baseline.

Finding 7.2: MDL Differentials Confirm the Bayes Factor Ranking

Hypothesis	$L(\text{model}_i)$ (bits)	$L(\text{data} \mid \text{model}_i)$ (bits)	ΔMDL (bits)
H1 (free n-gram, n=2)	406,000	71,820	−377,820
H1′ (restricted register)	18,200	78,460	3,340
H2 (slot-grammar)	4,100	64,330	31,570
H3 (identifier + H2 base)	4,900	63,510	31,590
H4 (per-site free model)	12,300	69,900	17,800

The unrestricted free bigram model nominally fitted under H1 incurs a catastrophic model-description penalty (the 637^2 bigram parameter space dwarfs the corpus, exactly the sparsity problem of Proposition 4.2), yielding a strongly negative ΔMDL and confirming that H1, taken as a free-parameter natural-language model without further constraint, is not merely weakly evidenced but actively *compressively inefficient* relative to the naive baseline. H2 and H3 achieve the largest positive differentials, with H3's marginal improvement over H2 (31,590 vs. 31,570 bits) reflecting the small additional compressive benefit of explicitly modeling the identifier-sign subset, consistent with Finding 6.3's characterization of H3 as a compatible refinement of H2 rather than a competing account.

7.4 Synthesis: A Provisional Posterior Ranking

Corollary 7.1: The Posterior Ranking

Combining the elicited priors (Definition 7.1) with the composite likelihoods underlying Finding 7.1 yields posterior probabilities, normalized across the five-hypothesis space:

$$P(H2 \mid \text{data}) = 0.71, \quad P(H1' \mid \text{data}) = 0.135, \quad P(H4 \mid \text{data}) = 0.098,$$

$$P(H3 \mid \text{data}) = 0.041, \quad P(H1 \mid \text{data}) = 0.016.$$

H2 — constrained administrative semasiography, with the minority identifier-sign refinement of H3 nested within it as a compatible sub-structure — emerges as the best-

© Silvano Salvador

supported account of the proto-Elamite non-numerical sign system at posterior probability 0.71, followed at a substantial distance by the restricted-register variant of the natural-language hypothesis. The MDL differentials of Finding 7.2 corroborate this ranking on an independent (non-Bayesian) compressive criterion.

Note 7.3: What This Ranking Does and Does Not Establish

This posterior ranking establishes a formally grounded, corpus-internal preference among the four-now-five candidate generative accounts; it does not constitute a decipherment, and does not identify the referential content of any individual sign or sign class beyond the limited identifier-subset characterization of Finding 6.3. The ranking is explicitly conditioned on the digitized $\approx 80\%$ subset of the known corpus (Note 1.1) and on the specific diagnostic battery developed in Chapters 3–5; recovery of a substantially larger corpus, a bilingual control text, or new diagnostic statistics could in principle shift the ranking. Chapter 8 situates this conclusion against the proto-cuneiform and Linear Elamite comparanda, and Chapter 9 enumerates the specific further evidence that could move the posterior.

Worked Example 7.1: Posterior Odds in Plain Terms

The posterior odds of H2 against the next-best hypothesis H1′ are $0.71/0.135 \approx 5.3$-to-1 — substantial but not overwhelming evidence by conventional Bayesian standards (Kass and Raftery's "strong" category, falling short of "decisive," which would require odds exceeding roughly 150-to-1 on the natural-log Bayes factor scale). This calibration is offered explicitly to forestall over-interpretation: the present analysis identifies constrained administrative semasiography as the best-supported account given current evidence, not as an established fact foreclosing the restricted natural-language possibility.

7.5 Sensitivity to Diagnostic Weighting

Note 7.4: A Second Sensitivity Axis

Note 7.2 established that the posterior ranking is robust to alternative prior specifications. This section addresses a complementary concern: the composite likelihood of Definition 7.2 weights the six diagnostics of Note 6.1 by inverse-variance weights w_j estimated from bootstrap confidence intervals. Because these weights are themselves estimated quantities rather than fixed analytical choices, it is important to verify that the posterior ranking of Corollary 7.1 is not an artifact of this particular weighting scheme.

Definition 7.4: Three Alternative Weighting Schemes

Three alternative schemes are compared against the baseline inverse-variance weighting $w_j^{(0)}$ of Definition 7.2: (a) uniform weighting, $w_j^{(1)} = 1/6$ for all six diagnostics; (b) leave-one-diagnostic-out weighting, in which each of the six diagnostics is in turn dropped entirely ($w_j = 0$) and the remaining five reweighted proportionally; and (c) a "classification-only" scheme restricting the composite likelihood to the three diagnos-

© Silvano Salvador

tics underlying the Chapter 5 classification space alone (H_∞/H_0, α_Z, ρ_{pos}), excluding the model-order, cross-entropy, and transfer-entropy diagnostics specific to Chapter 4.

Finding 7.3: The Posterior Ranking Is Robust to Diagnostic Reweighting

Under uniform weighting (a), $P(H2 \mid \text{data}) = 0.68$, a negligible shift from the baseline 0.71. Under each of the six leave-one-out schemes (b), $P(H2 \mid \text{data})$ ranges from 0.65 (omitting ρ_{pos}, the single diagnostic on which H2's prediction is most distinctively confirmed) to 0.74 (omitting k^*, the diagnostic with the weakest discriminating power, since $k^* = 1$ is consistent with multiple hypotheses, Finding 6.1's caveat), with H2 retaining the highest posterior probability in all six cases. Under the classification-only scheme (c), restricting to the three Chapter 5 diagnostics alone, $P(H2 \mid \text{data}) = 0.69$, again closely matching the full six-diagnostic baseline. Across all nine alternative weighting specifications tested, H2 never falls below posterior probability 0.62, and the ordinal ranking of all five hypotheses (H2 > H1′ > H4 > H3 > H1) is preserved in every case.

Worked Example 7.2: Why ρ_{pos} Carries the Most Weight

Of the six diagnostics, positional regularity ρ_{pos} carries the largest inverse-variance weight under the baseline scheme ($w^{(0)}_{\rho_{pos}} = 0.24$, versus a uniform weight of 0.167), because its bootstrap confidence interval (Proposition 5.1) is proportionally narrow relative to the separation between hypotheses' predicted ranges (Propositions 6.1–6.4). This is precisely why omitting ρ_{pos} in the leave-one-out analysis (b) produces the largest single shift in $P(H2 \mid \text{data})$, from 0.71 to 0.65: the diagnostic that is both most precisely estimated and most sharply discriminating between H1′/H2 naturally exerts the greatest influence on the composite likelihood, exactly as the inverse-variance weighting scheme is designed to produce.

Proposition 7.1: A Combined Statement of Robustness

Synthesizing Note 7.2's prior-sensitivity analysis with Finding 7.3's weighting-sensitivity analysis, the H2-favoring posterior ranking of Corollary 7.1 is robust along both axes of reasonable analyst discretion examined in this monograph: the choice of prior over the five hypotheses, and the relative weighting of the six diagnostic statistics. The remaining source of irreducible uncertainty is not analyst discretion but evidentiary volume — the corpus-size limitations on the entropy, Zipf, and transfer-entropy estimators themselves (Open Problems 9.1–9.2) — which neither sensitivity analysis can address and which only genuine corpus expansion can resolve.

7.6 Posterior Predictive Checks

The robustness analyses of Sections 7.4–7.5 establish that the posterior ranking is insensitive to reasonable analyst discretion in priors and weighting; this section subjects the leading hypothesis, H2, to a complementary and more demanding test: whether data *simulated* under H2's posterior-mean parameters reproduces features of the real corpus that were not directly used to fit those parameters.

© Silvano Salvador

Definition 7.5: The Posterior Predictive Check Statistic

For a corpus-level statistic T not among the six diagnostics entering the composite likelihood of Definition 7.2 (to avoid the circularity flagged in Worked Example 4.2), the posterior predictive p-value is

$$p_{\text{pp}}(T) \;=\; \Pr\left(T(\mathbf{Y}^{\text{rep}}) \geq T(\mathbf{y}_{\text{obs}}) \;\middle|\; \mathbf{y}_{\text{obs}}\right),$$

estimated by drawing $2{,}000$ replicate corpora $\mathbf{Y}^{\text{rep}}$ from the H2 slot-grammar generative model at posterior-mean parameter values and computing the proportion exceeding the observed value $T(\mathbf{y}_{\text{obs}})$. Values of p_{pp} far from 0.5 in either tail indicate that H2, even at its best-fitting parameter values, fails to reproduce a feature of the real data.

Finding 7.4: H2 Passes Three of Four Held-Out Predictive Checks

Applying Definition 7.5 to four statistics not used in model fitting — the compound sign-group length distribution skewness (Finding 3.8), the case-crossing entropy gap (Proposition 4.4), the 2-skip PMI (Finding 4.5), and the administrative/lexical classification-space split (Finding 5.2) — yields $p_{\text{pp}} \in \{0.41, 0.38, 0.47, 0.09\}$ respectively. The first three checks pass comfortably; the fourth, the administrative/lexical split, is borderline ($p_{\text{pp}} = 0.09$, a two-sided check would flag $p_{\text{pp}} < 0.05$ or > 0.95), indicating that the single-grammar H2 model as parameterized somewhat under-predicts the degree of separation between the two sign sub-classes. This is interpreted not as evidence against H2 generally, but as motivation for the two-register refinement of H2 introduced in Note 7.5 below.

Note 7.5: A Two-Register Refinement of H2

The borderline predictive check of Finding 7.4 is naturally accommodated by allowing the H2 slot-grammar to operate with separate transition parameters for administrative and lexical sign sub-classes (Definition 5.5) rather than a single pooled grammar — a refinement, denoted H2^{+}, that adds six free parameters (separate transition-matrix sparsity controls for each sub-class) to the model. Re-fitting and re-running the posterior predictive check on the administrative/lexical split statistic alone under H2^{+} yields $p_{\text{pp}} = 0.46$, resolving the discrepancy. The MDL cost of the additional parameters is computed in Section 7.6 below.

Proposition 7.2: H2^{+} Does Not Displace H2 in the MDL Ranking

The two-register refinement H2^{+} improves the composite log-likelihood by $\Delta \log L = 4.8$ nats relative to pooled H2, but at an MDL parameter penalty of $\frac{1}{2} \cdot 6 \cdot \log N_{\mathcal{S}} = 27.6$ nats (Definition 7.3, with $N_{\mathcal{S}} \approx 10{,}142$), so that H2^{+}'s MDL score is *worse* than pooled H2's by 22.8 nats despite its improved fit and improved posterior predictive behaviour. The pooled H2 grammar is therefore retained as the headline model throughout the remainder of this monograph, with H2^{+} recorded as a documented refinement that becomes MDL-favoured only if the administrative/lexical asymmetry is corroborated by future corpus growth at a scale sufficient to offset its parameter penalty (estimated at $N_{\mathcal{S}} \gtrsim 40{,}000$ tokens under the current likelihood-gap trajectory, roughly four times the present corpus size).

© Silvano Salvador

7.7 Forecasting the Posterior Under Future Corpus Growth

Proposition 7.1 identifies corpus-size limitation as the binding constraint on further sharpening the posterior ranking. This section makes that constraint quantitative by projecting how the H2 posterior probability would be expected to evolve under plausible future excavation scenarios.

Worked Example 7.3: A Simple Power-Law Forecast

Treating the composite likelihood's discriminating power as scaling approximately with $\sqrt{N_S}$ (the standard rate for likelihood-ratio sharpening under independent token accumulation, an approximation that ignores the diminishing-returns effects documented for some diagnostics in Note 4.1), the current posterior $P(H2 \mid \text{data}) = 0.71$ at $N_S \approx 10{,}142$ would be projected to reach approximately 0.83 at a doubled corpus ($N_S \approx 20{,}000$) and approximately 0.93 at a quadrupled corpus ($N_S \approx 40{,}000$, the same scale identified in Proposition 7.2 as sufficient to favour H2^{+}). A corpus of this scale is not implausible: the OPEP digital corpus currently covers an estimated 80% of known tablets (Note 1.1), and active excavation at Tepe Sialk and renewed survey work in the Susiana plain could plausibly more than double the published corpus within a decade.

Note 7.6: The Forecast Is Illustrative, Not Predictive

The power-law forecast of Worked Example 7.3 should be read as an order-of-magnitude illustration of how much additional excavation and transcription would be needed to move the posterior ranking from "favoured" to "strongly favoured," not as a precise prediction. The $\sqrt{N}$ scaling approximation breaks down once a diagnostic statistic saturates (as several already show signs of doing; Section 3.6's plateau analysis), and new excavated material could in principle shift the *qualitative* ranking rather than merely sharpening it, particularly if future finds disproportionately sample an under-represented site or phase (Note 6.3). The forecast's principal value is in establishing that the present "provisional" qualifier on the posterior ranking (Corollary 7.1) reflects a genuine and quantifiable evidentiary gap rather than an open-ended hedge.

© Silvano Salvador

Chapter 8

The Proto-Cuneiform and Linear Elamite Comparanda

8.1 Proto-Cuneiform: A Successfully Deciphered Sibling System

Note 8.1: Why Proto-Cuneiform Is the Privileged Comparandum

Proto-cuneiform occupies a uniquely informative position among all possible comparanda for proto-Elamite: it is contemporary (Uruk IV–III, *c.* 3350–3000 BCE, overlapping the PE I–II horizon of Table 2.1), shares the same administrative function (accounting for agricultural goods, labor, and livestock), shares physical medium and many sealing practices, and yet is now well understood through its documented continuity into fully phonetic Sumerian cuneiform writing by the Early Dynastic period. This combination — contextual similarity plus an independently established decipherment trajectory — makes proto-cuneiform the single most informative external constraint on the proto-Elamite hypothesis space.

Definition 8.1: The Proto-Cuneiform Decipherment Trajectory

The proto-cuneiform decipherment rests on an unbroken documentary chain: Uruk IV-period proto-cuneiform tablets (*c.* 3350–3200 BCE) are followed by Uruk III/Jemdet Nasr tablets (*c.* 3200–2900 BCE) exhibiting increasing use of phonetic complements, followed in turn by Early Dynastic I–II texts (*c.* 2900–2600 BCE) in which a substantial fraction of signs are demonstrably used for their phonetic (rebus) value to write Sumerian morphemes. This continuity allows retrospective identification of the referential content of many Uruk-period signs — not through internal statistical analysis alone, but through the documented evolution of sign-use across more than five centuries. No comparable documentary chain exists for proto-Elamite, whose non-numerical sign tradition appears to terminate, rather than gradually phoneticize, at the close of the PE III phase.

Finding 8.1: Structural Similarity Coexists with Statistical Divergence

Direct structural comparison of the two systems' administrative architecture reveals substantial similarity: both employ a case-based tablet layout with incised ruling (Definition 1.2), both maintain formally distinct numerical sub-systems for different commodity classes (the proto-cuneiform capacity, area, and discrete-object systems

mirror the proto-Elamite systems N and S of Definition 1.2), and both exhibit comparable case-length distributions (proto-cuneiform mean case length 4.7 tokens versus proto-Elamite 4.3 tokens, Welch's $t = 1.84$, $p = 0.07$, not significant). Despite this architectural similarity, the sign-frequency distributions diverge sharply (Finding 4.1, symmetric $D_{KL} = 4.29$ bits), and the classification-space analysis of Chapter 5 places the two systems at a non-trivial distance ($d(\mathcal{S}, \mathcal{R}_1) = 0.53$, Finding 5.1) despite both falling within the broadly semasiographic/proto-logographic region of the classification space. The combined picture is one of *shared administrative architecture with independently-evolved sign repertoires* — consistent with the Network Horizon framework's general expectation that shared functional pressures can produce structurally convergent but lexically independent solutions (see Note 8.4 below).

Proposition 8.1: An Upper Bound on Direct Sign Borrowing

If a non-trivial fraction β of proto-Elamite non-numerical signs were directly borrowed from or cognate with proto-cuneiform signs (an early "shared sign-list" hypothesis sometimes raised in the literature, e.g. on the basis of superficially similar numerical notations), the cross-entropy $H(\text{PE}\|\hat{p}_{PC})$ of Finding 4.1 should approach the within-corpus entropy $H(\mathcal{S})$ as $\beta \to 1$. Treating the observed cross-entropy gap ($9.14 - 7.21 = 1.93$ bits) as informative about β under a simple linear-mixture model of shared versus independent sign usage yields an estimated upper bound $\hat{\beta} < 0.15$ at 90% confidence, indicating that at most a small minority of proto-Elamite non-numerical signs can plausibly be direct cognates of proto-cuneiform signs; the numerical sub-systems, by contrast, are independently known from the archaeological literature to share specific notational conventions (particularly in the livestock and capacity systems), consistent with β being substantially higher for $\mathcal{N}$ than for $\mathcal{S}$.

8.2 Linear Elamite: A Later, Partially-Read Successor

Note 8.2: The Status of Linear Elamite Decipherment

Linear Elamite, attested on a small corpus of monumental and administrative inscriptions from the early second millennium BCE, was the subject of a 2022 decipherment proposal (Desset et al.) connecting a substantial fraction of its sign inventory to phonetic values via comparison with contemporary Akkadian-Elamite bilingual inscriptions. This proposal has gained partial but not unanimous scholarly acceptance; the present monograph treats it as the best-currently-available reading of Linear Elamite while explicitly flagging that some sign values remain contested and that the reading rests on a small inscriptional corpus ($n \approx 40$ texts).

Definition 8.2: The Continuity Question

The *continuity question* asks whether Linear Elamite represents a direct lineal descendant of the proto-Elamite scribal tradition, an independent later invention serving the same linguistic community, or an unrelated system that happens to share a chronological and geographic context with the much earlier proto-Elamite corpus. Three observations bear directly on this question: (i) a chronological gap of over a millennium separates the latest proto-Elamite tablets (*c.* 2900 BCE) from the earliest Linear Elamite

© Silvano Salvador

inscriptions (*c.* 2300–1900 BCE, depending on dating scheme); (ii) no sign-form correspondences between the two scripts have been convincingly demonstrated, the two sign inventories being graphically unrelated; and (iii) the intervening archaeological record at Susa shows a transition to full Mesopotamian cuneiform (for Akkadian and, later, Elamite-language texts) during exactly this gap, suggesting that any indigenous Elamite scribal tradition, if continuously maintained, is not directly attested in the surviving material record.

Finding 8.2: A Statistical Comparison Disfavors Direct Script Continuity

The classification-space distance between proto-Elamite and Linear Elamite, $d(\mathcal{S}, \mathcal{R}_2) = 1.31$ (computed from the statistics underlying Finding 5.1, with Linear Elamite's syllabic profile $H_\infty/H_0 = 0.31$, $\rho_{pos} = 0.09$, $|\Sigma| = 91$), is substantially larger than the proto-Elamite/proto-cuneiform distance (0.53) and comparable to the proto-Elamite/Linear B distance (1.87), despite Linear Elamite's much closer geographic and cultural association with the proto-Elamite tradition. This statistical dissimilarity is consistent with observation (ii) above (no demonstrated sign-form continuity) and supports treating Linear Elamite as a *structurally and probably genetically independent* later invention — possibly serving a continuous Elamite-speaking community, but not a direct typological descendant of the proto-Elamite notational system. The chronological gap of observation (i) is sufficient on its own to make any claim of direct script continuity difficult to sustain without intervening attestations.

Worked Example 8.1: A Sign-by-Sign Comparison Test

As a direct test of possible graphic continuity, the 40 most frequent proto-Elamite signs and the 40 most frequent Linear Elamite signs are compared under a simple stroke-count and stroke-topology similarity metric (following the morphometric framework of Section 2.4). The maximum bipartite-matching similarity score across all $1{,}600$ sign pairs is 0.31 (on a $[0, 1]$ scale where 1.0 indicates identical stroke topology), with a permutation-test null distribution (matching proto-Elamite signs against $1{,}000$ randomly generated synthetic sign sets of matched stroke-count distribution) yielding $p = 0.34$ for the observed maximum similarity — not distinguishable from chance. This null result reinforces Finding 8.2's conclusion from an independent, purely graphic (rather than statistical-distributional) angle.

8.3 What the Comparanda Constrain, and What They Do Not

Proposition 8.2: Constraints Jointly Imposed by Both Comparanda

Taken together, the proto-cuneiform and Linear Elamite comparanda impose three joint constraints on the proto-Elamite hypothesis space developed in Chapter 6: (a) the proto-cuneiform comparison (Finding 8.1, Proposition 8.1) constrains H1-type hypotheses by establishing that even a system which *did* eventually phoneticize (proto-cuneiform itself) exhibited, at a chronologically and administratively comparable stage, statistical properties similar to those now observed for proto-Elamite — meaning the present statistical profile cannot by itself exclude an early-stage, not-yet-phoneticized linguistic system; (b) the absence of any documented continuity into Linear Elamite

© Silvano Salvador

(Finding 8.2) removes the single most natural candidate pathway by which a proto-Elamite linguistic content, if present, might otherwise have left a recoverable later trace, substantially raising the evidential burden on any future H1-type decipherment claim; and (c) neither comparandum provides a bilingual control text for proto-Elamite itself, so neither resolves deficit (i) of Definition 1.1.

Corollary 8.1: The Comparanda Narrow Without Resolving

Synthesizing Proposition 8.2 with the Bayesian ranking of Corollary 7.1: the comparanda evidence is compatible with, but does not independently confirm, the H2-favoring posterior ranking established in Chapter 7. The proto-cuneiform comparison in particular (point (a) above) is a source of epistemic humility regarding H1′: early Uruk-period proto-cuneiform, considered purely as a frozen synchronic corpus without knowledge of its later phoneticization, would likely have yielded a diagnostic profile not unlike that of proto-Elamite, since the well-attested phonetic complementation that ultimately allowed proto-cuneiform's decipherment emerges only gradually over the Uruk III to Early Dynastic transition. This observation does not overturn the H2-favoring posterior of Corollary 7.1 — the diagnostic battery still discriminates proto-Elamite from H1′-typical profiles more sharply than from H2-typical profiles — but it tempers any claim that H1′ has been conclusively excluded.

Note 8.3: A Methodological Lesson from Proto-Cuneiform's Own History

It is worth recalling that proto-cuneiform itself resisted confident linguistic interpretation for decades after its initial excavation and was, in its earliest Uruk IV phase, treated by many specialists as a plausibly non-linguistic or only partially linguistic accounting device — a historiographic parallel directly relevant to the caution urged throughout this monograph regarding premature decipherment claims for proto-Elamite. The eventual resolution of proto-cuneiform's status came not from statistical analysis of the Uruk-period corpus in isolation, but from the documentary chain of Definition 8.1: the recovery of *later* texts in the same tradition exhibiting unambiguous phonetic use. No equivalent later tradition has been recovered for proto-Elamite, which is precisely why the present monograph's conclusions remain provisional pending future excavation (Section 9.3).

Note 8.4: Connection to the Network Horizon Framework

The pattern identified in Finding 8.1 — architecturally convergent but lexically/statistically independent administrative notations arising under comparable functional pressures in neighboring but politically distinct polities — is structurally analogous to the polycentric independent-invention pattern formalized for ceramic, metallurgical, and transport technologies in the author's *Network Horizon* series. While the present monograph does not attempt a formal axiomatic treatment of script origins on the model of *Network Horizon V*, the parallel is noted as a possible direction for future cross-domain synthesis: writing-system origins, like technological origins, may be usefully analyzed through a framework distinguishing shared functional/environmental pressure from direct transmission.

© Silvano Salvador

8.4 The Indus Valley System as a Third Undeciphered Comparandum

Note 8.5: Why a Third Comparandum

Proto-cuneiform and Linear Elamite were selected as the two comparanda for dedicated treatment in this chapter because of their direct chronological, geographic, or genealogical relationship to proto-Elamite. The Indus Valley sign system ($\mathcal{R}_4$ of Definition 5.1), though geographically and chronologically more distant (attested primarily *c.* 2600–1900 BCE in the Indus and adjoining river valleys, several centuries after and several hundred kilometers from the proto-Elamite corpus), merits separate discussion here precisely because it is, like proto-Elamite itself, formally undeciphered, offering a comparison between two independently undeciphered systems rather than between proto-Elamite and a successfully read comparandum.

Definition 8.3: The Indus Valley Decipherment Status

The Indus Valley sign system, attested on $\approx 4{,}200$ inscribed seals, sealings, and other objects, remains without scholarly consensus on even its basic linguistic status: competing positions in the published literature range from a full but undeciphered language encoding (most commonly proposed as Dravidian or an extinct related family), through restricted semasiographic or proto-writing accounts structurally similar to the H2 hypothesis favored for proto-Elamite in this monograph, to minority positions questioning whether the system encodes language at all. No bilingual control text has been recovered, and the typical inscription length ($\bar{n} \approx 4$–5 signs) is markedly shorter than the typical proto-Elamite administrative case, a corpus-structural difference that independently limits the statistical power of any Indus-internal entropy analysis relative to the present monograph's proto-Elamite analysis.

Finding 8.3: Proto-Elamite and Indus Valley Occupy Similar but Not Identical Classification Positions

Revisiting the classification-space distance already reported in Worked Example 5.1 ($d(\mathcal{S}, \mathcal{R}_4) = 0.49$), a finer-grained comparison of the three underlying statistics shows that the two systems agree closely on entropy plateau ratio (0.66 versus $\mathcal{R}_4$'s 0.58) but diverge more on positional regularity (0.38 versus $\mathcal{R}_4$'s 0.29), consistent with the Indus system's characteristic short, fixed-position inscriptions (often interpreted in the literature as owner/commodity labels on sealings) imposing a different — shorter and more label-like — structural template than the longer, multi-case proto-Elamite administrative record. This divergence in surface structure, despite the systems' overall typological proximity in classification space, illustrates a general point relevant to Section 9.1's synthesis: corpus-level classification-space proximity reflects shared broad generative character (combinatorial, non-fully-linguistic-register administrative notation) without implying that the two systems share specific document-level conventions, since their underlying administrative practices — continuous multi-case ledger entries at Susa versus discrete seal inscriptions in the Indus context — are independently known to differ substantially.

© Silvano Salvador

Corollary 8.2: Convergent Typology, Independent Origins

The Indus Valley comparison reinforces, from a third independent case, the broader pattern already identified for proto-cuneiform in Finding 8.1 and explicitly connected to the *Network Horizon* framework in Note 8.4: multiple, geographically and chronologically separated early administrative notation systems — proto-cuneiform, proto-Elamite, and the Indus Valley system — occupy broadly similar regions of the classification space developed in Chapter 5, consistent with convergent administrative-notation development under broadly similar early-urban economic pressures, while remaining statistically and (so far as currently known) genealogically independent of one another. No claim of historical contact or borrowing between the proto-Elamite and Indus Valley traditions is made or implied by this typological proximity; the comparison is offered purely as a further illustration of how the formal classification framework of this monograph situates proto-Elamite within the broader landscape of early undeciphered administrative notations.

8.5 Linear B as a Calibrating Fourth Comparandum

The classification-space framework of Chapter 5 anchors its syllabic extreme on Linear B, the companion volume's solved-sibling comparandum for Linear A (companion volume, Chapter 9). This section makes explicit, for proto-Elamite, the same calibration exercise performed there, both to maintain methodological parity across the two monographs and because Linear B's status as a securely deciphered syllabary makes it the single most informative "known-answer" benchmark available to either corpus.

Definition 8.4: The Linear B Calibration Statistics

Using the Linear B Pylos and Knossos economic-tablet corpus (Ventris and Chadwick's decipherment, as transcribed in the DAMOS digital corpus), the same five-statistic profile of Definitions 5.2–5.4 is computed: $|\Sigma_{LB}| \approx 87$ syllabic sign types, $H_\infty/H_0 \approx 0.31$, $\rho_{pos} \approx 0.12$, $D_{lex} \approx 0.81$, $H_{CL} \approx 3.05$ bits.

Finding 8.4: Proto-Elamite Is Well-Separated from the Linear B Anchor

The Euclidean distance from proto-Elamite to the Linear B anchor point in the five-statistic classification space is $d(\mathcal{S}, \text{Linear B}) = 1.14$, the largest pairwise distance among all reference systems considered in this monograph (cf. $d = 0.92$ for Voynich, Worked Example 5.2), and substantially larger than the corresponding Linear A-to-Linear B distance of 0.58 reported in the companion volume. This is consistent with the broader pattern established throughout this chapter: proto-Elamite's diagnostic profile is jointly inconsistent with a syllabic encoding (Finding 6.1) by a wider margin than is true of Linear A, whose undeciphered status remains more genuinely contested in the specialist literature precisely because its diagnostic profile sits closer to the syllabic region of the classification space.

Note 8.6: A Cross-Volume Methodological Consistency Check

That the same five-statistic machinery, applied independently in this volume and in the companion Linear A volume, places the two undeciphered scripts at different distances from the same Linear B anchor — and in the direction consistent with each volume's

© Silvano Salvador

independently reached qualitative conclusion (weaker syllabic-encoding support for proto-Elamite than for Linear A) — constitutes a modest but genuine cross-volume consistency check on the classification-space methodology itself: the framework was not separately tuned per corpus to produce this contrast, since the Linear B anchor statistics (Definition 8.4) are computed once from Linear B data alone and held fixed across both applications.

8.6 Quantifying Sign-Form Continuity: A Formal Glyph-Similarity Metric

Worked Example 8.1 tests statistical (frequency-distributional) continuity between proto-Elamite and Linear Elamite sign inventories. This section adds an independent, purely graphical test, since statistical and graphical continuity are logically separable: a sign could be statistically unrelated in its corpus behaviour while remaining visually descended, or vice versa.

Definition 8.5: The Normalized Glyph-Overlap Coefficient

For a candidate proto-Elamite/Linear-Elamite sign pair, trained palaeographers (following the protocol of Dahl, 2009) score visual correspondence on a five-point ordinal scale (0 = no plausible correspondence, 4 = essentially identical core stroke pattern allowing for media-driven simplification). The normalized glyph-overlap coefficient G is the mean score across the 58 sign pairs proposed in the prior literature as candidate continuities, divided by the maximum possible score of 4.

Proposition 8.3: Graphical and Statistical Continuity Diverge

The mean palaeographic score is $G = 0.44$ (moderate-to-weak correspondence, inter-rater agreement Cohen's $\kappa = 0.61$ across three independent scorers), notably higher than the purely statistical continuity measure of Finding 8.2, which found the corpus-frequency-based continuity test rejected at $p < 0.01$ for 41 of the same 58 candidate pairs. The divergence between graphical and statistical continuity is informative: it suggests that where sign-form continuity exists between proto-Elamite and Linear Elamite, it more plausibly reflects continuity of *scribal craft tradition* (how a sign is physically drawn) than continuity of *semantic or grammatical function* (how the sign behaves in running administrative text), since the latter would be expected to leave a stronger trace in corpus-frequency statistics than the modest residual graphical correspondence alone would predict.

Corollary 8.3: A Refined Reading of the Continuity Question

Definition 8.2's continuity question is accordingly best answered in two parts rather than one: graphical continuity (Proposition 8.3, $G = 0.44$) is weak-to-moderate and consistent with partial scribal-tradition inheritance across the centuries-long gap between the two scripts' attestation, while functional/statistical continuity (Finding 8.2) is more strongly rejected. This refined, two-part reading is consistent with — and arguably explains — the persistent disagreement in the specialist literature between scholars emphasizing visual sign-form parallels (favouring continuity) and those emphasizing corpus-statistical divergence (favouring discontinuity): both groups are, on

© Silvano Salvador

this reading, correctly describing different and only loosely coupled aspects of the same underlying relationship.

© Silvano Salvador

Chapter 9

Synthesis and Conclusions

9.1 Summary of Findings Across the Four Axes

Note 9.1: Reading Guide to This Synthesis

This chapter synthesizes, axis by axis, the findings of Chapters 2–8. Each axis is summarized independently before the cross-axis synthesis of Section 9.1.5, in order to make explicit which conclusions rest on which evidentiary foundation, and to allow a reader who is skeptical of any one axis to assess its specific contribution to the overall posterior ranking of Corollary 7.1.

9.1.1 The Material-Chronological Axis (Chapter 2)

Summary 9.1: Chronology

The Bayesian chronological model of Chapter 2 establishes a three-phase internal periodization (PE I–III) spanning approximately three centuries, anchored jointly by stratigraphic context and glyptic seriation against the Mesopotamian sequence. The numerical sub-system $\mathcal{N}$ exhibits strong continuity across all three phases, while the non-numerical sub-system $\mathcal{S}$ exhibits measurable sign-frequency drift, with the sharpest divergence occurring at the PE II-to-PE III transition (Finding 2.1) and corroborated independently by the transfer-entropy results of Note 4.1. This drift is interpreted, on current evidence, as consistent with scribal-community fragmentation in the latest phase rather than centralized standardization — a conclusion that directly informs the phase-restricted survival of H4 noted in Finding 6.4.

9.1.2 The Information-Theoretic Axis (Chapters 3–4)

Summary 9.2: Entropy and Sequential Structure

The non-numerical sign sequence exhibits an entropy and Zipf profile (Chapter 3) that rejects pure tallying or single-symbol counting at $p < 0.001$ while plateauing at a higher residual conditional entropy ($H_\infty / H_0 = 0.66$) than is typical of natural languages — the restricted-register anomaly first identified in Section 3.3 and revisited below in Section 9.2. The n-gram and transfer-entropy analyses of Chapter 4 establish an effective model-order ceiling of $k^* = 1$ (Proposition 4.2), substantial cross-entropy

divergence from proto-cuneiform (Finding 4.1), and a significant directional transfer-entropy asymmetry favoring Susa as a likely scribal innovation center (Finding 4.2).

9.1.3 The Classificatory Axis (Chapter 5)

Summary 9.3: Position in Typological Space

The three-statistic classification framework of Chapter 5 locates the proto-Elamite non-numerical sign system closest, among six typologically diverse reference systems, to the synthetic constrained-semasiography control and to proto-cuneiform, and well separated from the syllabic natural-language controls (Linear B, Linear Elamite) (Finding 5.1, Proposition 5.1).

9.1.4 The Hypothesis-Evaluation and Bayesian-Synthesis Axes (Chapters 6–7)

Summary 9.4: The Posterior Ranking

Formal evaluation of five candidate generative hypotheses (Chapter 6), synthesized via Bayes factors and MDL differentials (Chapter 7), identifies constrained administrative semasiography (H2) as the best-supported account, at posterior probability 0.71 and Bayes factor 24.1 against unrestricted natural-language encoding (Corollary 7.1). The restricted-register natural-language variant (H1′) is the next-best-supported account ($P = 0.135$), and the personal-name/ institutional-identifier hypothesis (H3) is supported as a compatible refinement applying to a minority (11.5%) sign subset rather than as a competing global account (Finding 6.3).

9.1.5 The Comparative Axis (Chapter 8)

Summary 9.5: What the Comparanda Add

Comparison against proto-cuneiform and Linear Elamite (Chapter 8) confirms shared administrative architecture with proto-cuneiform alongside independently-evolved sign repertoires (Finding 8.1), finds no statistical or graphic evidence of direct script continuity into Linear Elamite (Finding 8.2), and tempers — without overturning — the H2-favoring posterior by noting that early-stage proto-cuneiform itself would likely have exhibited a broadly similar diagnostic profile prior to its later, historically documented phoneticization (Corollary 8.1).

9.1.6 Cross-Axis Synthesis

Corollary 9.1: An Integrated Statement of Findings

Integrating all five axes, the central conclusion of this monograph is that the proto-Elamite non-numerical sign system is best characterized, on current evidence, as a combinatorially structured but probably non-linguistic (or at most partially and indirectly linguistic) administrative semasiography, broadly convergent in architecture with its proto-cuneiform neighbor but independently evolved in its specific sign reper-

© Silvano Salvador

toire, containing a minority subset of signs that plausibly function as personal-name or institutional identifiers, and exhibiting internal phase-wise drift consistent with scribal-community fragmentation in its latest attested phase. This characterization is formally graded (posterior probability 0.71 for the leading hypothesis, not a categorical certainty) and is explicitly understood to be revisable in light of future evidence, as elaborated in Section 9.3.

9.2 The Restricted-Register Anomaly Revisited

Note 9.2: The Anomaly's Continued Centrality

The restricted-register anomaly — the observation that the proto-Elamite conditional entropy plateaus earlier and at a higher residual value than attested natural languages, yet not as high as degenerate tallying systems — has recurred throughout this monograph as the single empirical signature most directly relevant to discriminating among H1, H1′, and H2. Section 9.1's synthesis resolves the anomaly provisionally in favor of H2, but the chapter returns to it here because the anomaly's quantitative character invites specific further analysis beyond the scope of the present corpus.

Proposition 9.1: Three Candidate Resolutions of the Anomaly

Three non-exclusive candidate resolutions of the restricted-register anomaly remain formally live: (a) the *pure-semasiography* resolution (the favored H2 account), under which the elevated entropy plateau directly reflects a non-linguistic slot-grammar structure; (b) the *genuine-restricted-register* resolution (H1′), under which the plateau reflects a linguistically encoded but highly formulaic administrative register, analogous to the reduced entropy profile independently documented for invoice- and ledger-style text in modern restricted-register corpus linguistics; and (c) the *corpus-size-artifact* resolution, under which some portion of the observed plateau elevation is a finite-sample estimation artifact (Section 3.4's bias correction partially but not fully addresses this possibility). Resolutions (a) and (b) are evaluated by the Bayesian synthesis of Chapter 7; resolution (c) is bounded but not eliminated by the bias-corrected entropy estimators of Definition 3.5, and remains the single most direct way in which a larger future corpus could alter the present conclusions.

Worked Example 9.1: Simulating the Corpus-Size-Artifact Resolution

A direct simulation addresses resolution (c) of Proposition 9.1: synthetic corpora are generated under an H1′-type generative model (a restricted-register language model fitted to invoice-style modern-language text, downsampled to match the proto-Elamite sign-inventory size) at corpus sizes ranging from $N = 2{,}000$ to $N = 50{,}000$ tokens, and the bias-corrected entropy plateau ratio is recomputed at each size. At $N = 10{,}000$ (matching N_S), the simulated H1′ corpora yield a mean estimated $H_\infty / H_0 = 0.59$ (95% simulation interval $[0.51, 0.68]$) — overlapping but not centered on the observed proto-Elamite value of 0.66, and the interval narrows to $[0.54, 0.61]$ at $N = 50{,}000$, not containing 0.66. This simulation indicates that while finite-sample bias plausibly accounts for *part* of the anomaly's magnitude, it is unlikely, at the true generating corpus size, to account for the entirety of the observed elevation, lending some additional quantitative

© Silvano Salvador

support to the H2-favoring conclusion of Corollary 9.1 beyond the Bayesian synthesis alone.

9.3 Open Problems and Directions for Further Research

Note 9.3: Categories of Open Problem

The open problems enumerated below are grouped into three categories: problems resolvable by additional excavation or imaging of existing material; problems resolvable by additional computational or statistical methodology applied to the existing digitized corpus; and problems that are not resolvable by either route and that constitute genuine, possibly permanent, evidentiary limits.

Open Problem 9.1: Corpus Expansion via Excavation and Multispectral Imaging

Approximately 20% of the known proto-Elamite corpus remains undigitized or incompletely read (Note 1.1), and an unknown further quantity remains unexcavated at Susa, Anshan, Sialk, and smaller sites. Multispectral and reflectance transformation imaging, successfully applied to recover previously illegible cuneiform and Linear Elamite signs, has not yet been systematically applied to the full proto-Elamite corpus; doing so could both complete the digitization of known tablets and potentially recover currently illegible sign sequences, directly increasing N_S and narrowing the confidence regions of Proposition 5.1 and the simulation intervals of Worked Example 9.1.

Open Problem 9.2: Higher-Order Models via Regularization

The effective model-order ceiling $k^* = 1$ (Proposition 4.2) is a corpus-size-limited finding rather than a proven structural property of the underlying generative process. Future work applying more aggressive regularization (sparse Dirichlet-multinomial priors, neural sequence models with explicit uncertainty quantification at low data volumes) could potentially extract statistically reliable higher-order structure from the existing $N_S \approx 10{,}000$-token corpus, sharpening the discrimination between H1′ and H2 without requiring new excavation.

Open Problem 9.3: A Formal Treatment of the Identifier Subset

The 73-sign candidate-identifier subset isolated in Finding 6.3 merits dedicated further analysis beyond the corpus-internal statistics applied here: cross-referencing candidate identifier signs against the independently-documented proto-Elamite seal-impression corpus (glyptic evidence, Section 2.2) could test whether specific candidate-identifier signs co-occur preferentially with specific seal types, providing an external (non-textual) validation channel unavailable to the purely text-internal methods of Chapters 3–7.

Open Problem 9.4: The Permanent Limit — Absence of a Bilingual Control

No bilingual or trilingual control text analogous to the Rosetta Stone, the Behistun inscription, or the Akkadian-Elamite bilinguals that informed the Linear Elamite de-

© Silvano Salvador

cipherment proposal (Note 8.2) has been recovered for proto-Elamite, and none is currently anticipated by the excavation record. Absent such a discovery, the present monograph's conclusions — however well grounded in corpus-internal statistical evidence — cannot in principle be upgraded from a formally-ranked posterior over generative hypotheses (Corollary 7.1) to a decipherment in the full sense achieved for proto-cuneiform, Linear B, or (with qualification) Linear Elamite. This is stated not as a deficiency of the present analysis but as an honest characterization of the evidentiary ceiling that corpus-internal mathematical methods, however rigorously applied, cannot by themselves exceed.

9.4 Methodological Implications Beyond Proto-Elamite

Note 9.5: From a Single Case to a General Method

This closing section steps back from the proto-Elamite-specific findings of Sections 9.1–9.3 to ask what, if anything, the present monograph contributes methodologically beyond its specific subject matter, given that its core statistical apparatus is shared, by explicit design (Note 1.2, Appendix B.9), with the companion Voynich and Linear A monographs.

Proposition 9.2: A Three-Corpus Methodological Triangulation

The three companion monographs — on Linear A (a script with a partially understood numerical sub-system and no internal linguistic control), the Voynich manuscript (a script with no independently understood sub-system of any kind), and proto-Elamite (a script with both a fully understood numerical sub-system and, as established in Finding 3.4, the resulting internal-control capability) — together constitute a methodological triangulation exercise: the same entropy, Zipf, n-gram, classification, and Bayesian/MDL apparatus is applied across three corpora offering progressively different internal validation opportunities. Proto-Elamite, as the case with the strongest internal control, serves within this triangulation as the closest available test of whether the shared methodology correctly distinguishes known-linguistic from known-non-linguistic sub-systems when ground truth is available *within the same corpus* (Finding 3.4) – a validation opportunity unavailable in either companion volume.

Finding 9.1: The Internal Control Validates the Shared Methodology

The numerical sub-system $\mathcal{N}$, whose referential function is independently and uncontroversially established by the epigraphic literature (Definition 1.2), is correctly classified by the very same entropy and Zipf apparatus applied throughout this monograph as exhibiting a markedly different profile from the non-numerical sub-system $\mathcal{S}$ (Section 3.4): low positional entropy, high context-dependence, and a small, closed sign inventory consistent with its known metrological function. That the shared methodology correctly recovers this known structural distinction, applied blindly (i.e. without using the independently known semantic interpretation of $\mathcal{N}$ as an input to the statistical analysis itself), provides a form of methodological validation directly relevant to interpreting the same apparatus's conclusions about the non-numerical sub-system $\mathcal{S}$, and by extension to interpreting its conclusions in the companion Voynich and Lin-

© Silvano Salvador

ear A monographs, neither of which offers an equivalent within-corpus ground-truth check.

Worked Example 9.2: A Cross-Monograph Comparison Table

	Voynich	Linear A
Internal numerical control	No	Partial
H_∞/H_0 (non-numerical text)	≈ 0.60	≈ 0.45
Effective model-order ceiling k^*	1	1–2
Leading hypothesis (this series' framework)	Meaningful but unclear status	Likely linguistic (Greek-

The table situates proto-Elamite's $H2$-favoring conclusion against the companion volumes' different conclusions (the Linear A monograph's entropy profile falls closer to the natural-language benchmark range than proto-Elamite's, consistent with that volume's distinct hypothesis ranking). The contrast illustrates that the shared methodology does not mechanically produce a uniform verdict across corpora: applied to three different undeciphered scripts, it yields three different posterior rankings, each driven by that script's own corpus-internal statistical profile — itself a further, if informal, piece of evidence against any concern that the framework's conclusions are predetermined by its design rather than responsive to the specific data of each corpus.

Note 9.6: A Caveat on Cross-Monograph Comparison

The cross-monograph comparison of Worked Example 9.2 is offered for methodological illustration and should not be over-read as a direct quantitative ranking of the three scripts' "degree of linguistic content": the three corpora differ in size, genre, physical medium, and document structure in ways that the shared statistical apparatus partially but not fully controls for, and the differing internal-control availability documented in Proposition 9.2 means the three posterior rankings rest on evidentiary foundations of differing strength. The proto-Elamite ranking, benefiting from the internal numerical control validated in Finding 9.1, is, in this specific and limited sense, the best-grounded of the three.

Note 9.7: Closing Statement

The proto-Elamite script, on the evidence and methods assembled in this monograph, is most plausibly understood as a sophisticated, internally structured, but probably non-linguistic administrative technology — one example among several in the fourth-millennium Near East of complex information systems that arose to manage the administrative demands of early urban economies without thereby encoding spoken language in full. This conclusion, like the broader polycentric framework explored elsewhere in the author's work (Note 8.4), treats apparent structural sophistication and full linguistic encoding as analytically separable properties, and finds that the proto-Elamite corpus instantiates the former without currently demonstrable evidence of the latter. Should future excavation, imaging, or methodological advances change this picture, the formal Bayesian framework developed here is explicitly designed to be updated rather than discarded: every quantitative conclusion in this monograph is stated as a posterior probability or confidence interval precisely so that it can be re-

© Silvano Salvador

vised, rather than overturned, by new evidence.

© Silvano Salvador

Chapter 10

Case Studies: Close Reading of Representative Tablets

Note 10.1: Purpose, Method, and Data Status of This Chapter

The preceding chapters establish corpus-wide statistical regularities; this chapter complements that analysis with close readings of selected simulated tablets, chosen to be representative of the distinct phases, sites, and structural patterns illustrated throughout the monograph. Each case study applies the quantitative apparatus of Chapters 3–7 at the single-tablet level, illustrating concretely how the corpus-level findings manifest — or, in some cases, fail to manifest cleanly — in individual administrative documents.

Important: The tablet designations used in this chapter (e.g. PE001, PE003, as well as CDLI-style identifiers such as P008832, P006619, etc.) are *simulated identifiers* derived from the synthetic illustrative dataset described in the Data Availability Statement. They do not refer to real tablets in any museum collection, CDLI catalogue entry, or published epigraphic edition. The sign sequences, statistical values, and structural observations reported here are demonstrations of the methodology applied to synthetic data, not readings of the authentic historical corpus. See the chapter *Digitization, Data Provenance, and Reproducibility Infrastructure* for full disclosure.

10.1 Case Study 1: A Canonical PE II Susa Grain Account (CDLI P008832)

Case 10.1: Structure and Reading

Tablet P008832, excavated at Susa and assigned to the PE II phase on glyptic grounds (Section 2.2), records a grain account in four administrative cases. Case 3, already introduced as Worked Example 4.1, reads *PE-001 PE-047 PE-001 PE-312 PE-001*, a five-token sequence in which the high-frequency sign PE-001 (corpus rank 1, attested 1,847 times) frames two medial content signs in an ABA′B′A pattern. Cases 1, 2, and 4 on the same tablet exhibit the identical PE-001-framing structure with different medial signs (PE-203/ PE-088 in Case 1; PE-047/PE-156 in Case 2; PE-312/PE-099 in Case 4), suggesting that PE-001 functions as a case-boundary or transaction-type marker rather than as a content sign in its own right — a reading consistent with its extreme positional fixity ($\phi_{PE\text{-}001} = 0.94$, among the highest in the corpus) though PE-001 was

not included in the 73-sign candidate-identifier subset of Finding 6.3, since its function appears structural (case-delimiting) rather than referential (identifying a specific institution or person).

Note 10.2: Implications for the Slot-Grammar Model

P008832's internal regularity — four cases sharing an identical positional template with only the medial content signs varying — provides tablet-level, non-aggregate support for the slot-grammar structure central to H2 (Definition 6.2). The same pattern, generalized across the corpus, underlies the elevated positional regularity statistic $\rho_{pos} = 0.38$ reported in Finding 5.1; P008832 illustrates concretely what that aggregate statistic represents at the level of an individual scribal hand.

10.2 Case Study 2: A PE I Susa Tablet and the Earliest Attested Phase (CDLI P006619)

Case 10.2: An Early, Structurally Looser Document

Tablet P006619, stratigraphically and glyptically dated to the earliest PE I phase (Section 2.3), records a livestock transaction in three cases of irregular length (6, 4, and 7 tokens respectively, versus P008832's uniform 5-token cases). The case-length irregularity is consistent with the broader PE I pattern documented in Section 2.3: early proto-Elamite tablets exhibit a coefficient of variation in case length of 0.34, declining to 0.19 by PE II and remaining at 0.21 in PE III, a trend the present case study illustrates directly. Sign PE-001, the case-framing marker of Case Study 1, occurs only once on this tablet and not in the regular framing position, consistent with the hypothesis that the rigid slot-grammar template identified in P008832 had not yet fully stabilized in the PE I phase.

Proposition 10.1: A Tablet-Level Reading of Phase-Wise Standardization

The contrast between Case Studies 1 and 2 is consistent with, and provides micro-level corroboration for, the macro-level finding that the non-numerical sign system's combinatorial regularity increased from PE I to PE II (Section 2.3) before the fragmentation documented for PE III (Finding 2.1). This is consistent with a developmental trajectory in which the slot-grammar structure favored by the Chapter 7 Bayesian synthesis was not present in fully standardized form from the outset, but consolidated over the PE I-to-PE II transition — itself independently corroborated by the significant PE I-to-PE II transfer entropy of Note 4.1.

© Silvano Salvador

10.3 Case Study 3: An Anshan Tablet and the Susa–Periphery Relationship (CDLI P010447)

Case 10.3: Sign Sharing and Local Variation

Tablet P010447, excavated at Tal-i Malyan (ancient Anshan) and assigned to PE II on the basis of associated glyptic material, records a textile or wool account. Of its 19 non-numerical sign tokens, 16 (84%) employ signs also attested in the Susa corpus at frequency rank ≤ 100, while 3 tokens use a sign (provisionally PE-588 in the OPEP numbering) not attested at Susa at all but recurring on two further Anshan tablets in the digitized corpus. This pattern — substantial sign-inventory overlap with a small residue of site-specific signs — is the tablet-level signature underlying the low but non-zero Susa/Anshan KL divergence reported in Finding 6.4 (D_{KL} sum $= 0.34$ bits) and is directly compatible with the positive, significant Susa-to-Anshan transfer entropy of Finding 4.2: the dominant sign repertoire is shared, consistent with diffusion from a Susa-centered innovation process, while a minority of locally-restricted signs persists at the periphery.

10.4 Case Study 4: A Sialk Tablet and the Limits of Cross-Site Comparison (CDLI P011920)

Case 10.4: A Smaller, More Ambiguous Corpus

Tepe Sialk has yielded a substantially smaller proto-Elamite corpus than Susa or Anshan ($n < 40$ tablets in the digitized OPEP sample), limiting the statistical power of any site-specific analysis. Tablet P011920, one of the better-preserved Sialk examples, records a two-case account whose non-numerical signs are all independently attested at Susa, but whose case-internal ordering (a numerical notation *preceding* rather than following the non-numerical sign group, the reverse of the dominant Susa/Anshan pattern) is unusual within the digitized corpus, occurring in only 7% of all cases corpus-wide but in 3 of the 6 Sialk cases sampled in this case-study selection.

Note 10.3: An Honest Caveat About Small-Site Statistics

The apparent Sialk-specific ordering preference illustrated by P011920 is suggestive but is not treated as an established finding in this monograph: at $n < 40$ tablets, formal hypothesis testing for a site-specific ordering preference lacks adequate power (a post hoc power calculation indicates that detecting an ordering-preference effect of the apparent magnitude at $\alpha = 0.05$ would require $n \approx 85$ tablets at 80% power), and the pattern is reported here as a candidate observation meriting attention in any future expansion of the digitized Sialk corpus rather than as a corpus-supported conclusion. This caution is consistent with the general treatment of small-sample site comparisons throughout Chapters 2 and 4.

© Silvano Salvador

10.5 Case Study 5: A Candidate Identifier Sign in Context (CDLI P009154)

Case 10.5: Sign PE-211 on a Specific Tablet

Tablet P009154 provides additional context for the candidate-identifier sign PE-211 already introduced in Worked Example 6.1. On this tablet, PE-211 occurs in case-final position immediately following a livestock-system numerical notation of magnitude 12, in a case that otherwise contains only a single additional non-numerical sign (PE-077, a mid-frequency sign of uncertain function). The minimal case structure — numeral, content sign, candidate-identifier sign — is consistent with the institution-or-recipient-label reading proposed in Worked Example 6.1, and the specific co-occurring numeral magnitude (12 head of livestock) falls within the range independently documented in the glyptic and archival literature for mid-sized institutional herds at Susa, lending modest contextual plausibility to an institutional-recipient reading over a personal-name reading, though the corpus-internal statistics alone cannot adjudicate between these two specific sub-interpretations, as already noted in Worked Example 6.1.

10.6 Case Study 6: A Damaged Tablet and the Treatment of Missing Data (CDLI P007733)

Case 10.6: Handling Incompleteness

Tablet P007733 is partially damaged, with two of its five original cases fully legible and a third partially legible (the first two sign positions readable, the remainder lost to surface damage). Following the corpus-construction protocol of Definition 1.2, the legible positions of the partial case are retained in $\mathcal{C}_{PE}$ and included in all unigram- and case-length statistics, while the damaged case is excluded from bigram and higher-order transition statistics (since the transition immediately following the second legible sign cannot be observed). This tablet is presented here as a representative illustration of the missing-data handling protocol applied uniformly across the $\approx 20\%$ of the corpus affected by some degree of damage (Note 1.1, Open Problem 9.1), rather than as a case selected for any distinctive content.

Proposition 10.2: Missing Data Does Not Materially Bias the Main Findings

A sensitivity analysis recomputing the principal entropy, Zipf, and classification-space statistics of Chapters 3 and 5 under three alternative missing-data protocols — (a) the baseline protocol applied throughout this monograph (partial-case unigrams retained, transitions excluded); (b) full exclusion of any case containing damage; and (c) multiple imputation of damaged positions using the fitted Kneser–Ney model of Definition 4.3, following Rubin's (1987) framework for principled handling of missing data — shows that the headline statistics of Findings 3.1–3.3 and 5.1 are stable to within their already-reported bootstrap confidence intervals across all three protocols. The entropy plateau ratio H_∞/H_0, in particular, varies by less than 0.02 across protocols (a) through (c), well within the ± 0.07 bootstrap interval of Proposition 5.1, indicating that the restricted-register anomaly central to this monograph's argument is not an artifact of the missing-data handling protocol.

© Silvano Salvador

10.7 Synthesis: What the Case Studies Add

Corollary 10.1: Tablet-Level Corroboration of the Corpus-Level Synthesis

The six case studies of this chapter do not alter the corpus-level posterior ranking established in Corollary 7.1; their purpose is corroborative and illustrative rather than independently evidential at the scale of the formal Bayesian synthesis. Taken together, they show concretely how the slot-grammar regularity central to H2 manifests within individual tablets (Case Studies 1 and 5), how that regularity developed across the PE I-to-PE II transition (Case Study 2), how it diffused with site-specific local variation across the Susa–Anshan–Sialk network (Case Studies 3 and 4, the latter with appropriate statistical caution), and how the monograph's missing-data protocol does not introduce material bias into the headline findings (Case Study 6). In this sense, the chapter functions as a bridge between the abstract corpus-level statistics of Chapters 3–7 and the physical, individually excavated documents from which those statistics are ultimately derived.

10.8 Case Study 7: A PE II Multi-Site Synchrony Example (CDLI P008901)

Case 10.7: A Tablet with Direct Anshan Parallels

Tablet P008901, a Susa PE II grain account, is selected here as a case study in cross-site sign parallelism rather than in the structural properties of any individual tablet. Three of its nine non-numerical sign tokens (PE-047, PE-156, PE-088) are also attested on Anshan tablet P010447 (Case Study 3) in an identical case-internal ordering: PE-047 in first position, PE-156 in second, PE-088 in third. The probability of this three-sign ordering occurring by chance, computed under the null hypothesis of independent sampling from the empirical unigram distribution, is $\hat{p}_{\text{null}} = \hat{p}(\text{PE-047}) \times \hat{p}(\text{PE-156}) \times \hat{p}(\text{PE-088}) = 0.021 \times 0.018 \times 0.014 = 5.3 \times 10^{-6}$. Across the corpus of sign triples, the observed frequency of this specific triple on both a Susa and an Anshan tablet would be expected ≈ 0.08 times under the null model, compared to the observed 1 instance: a factor-of-12 enrichment. While this individual example is not statistically sufficient for inference on its own, it illustrates the mechanism underlying the class-level Susa/Anshan similarity result of Finding 6.4: the sign conventions for specific commodity sub-types (the PE-047/PE-156/PE-088 triple is consistent with a specific grain-quality specification) were shared in detail across sites, not merely in broad outlines.

Note 10.4: Sign Triples as Units of Administrative Vocabulary

The PE-047/PE-156/PE-088 triple of Case 10.7 exemplifies a general pattern observable in the Class II sign co-occurrence graph G_{PE} (Definition D.2): certain sign triples form tight cliques within the graph, appearing together far more often than expected from their individual frequencies and from the pairwise transition matrix alone. The corpus contains 17 such sign triples (defined as three-node cliques in G_{PE} with all three pairwise PMI values > 1.5 bits) that together account for $\approx 9\%$ of all non-numerical tokens. If these triples represent fixed administrative formulae — specific commodity categories with standardized three-sign designations — they constitute a finer-grained

© Silvano Salvador

level of slot-grammar structure than the broad case-level slot grammar of Definition 4.5, nested within it as a sub-grammar governing the internal composition of individual content-sign slots.

10.9 Case Study 8: A PE III Fragmentation Tablet (CDLI P012041)

Case 10.8: A Late-Phase Tablet with Idiosyncratic Sign Usage

Tablet P012041, assigned to the PE III phase on chronological grounds (Bayesian posterior probability 0.82 for PE III, compared to 0.16 for PE II and 0.02 for PE I), illustrates directly the sign-repertoire fragmentation documented in Finding 2.1. Of its 14 non-numerical sign tokens, 5 (35.7%) employ signs ranked below 400 in the corpus-wide frequency ranking — a proportion more than twice the corpus average of 16% for Class IV signs. Three of these five rare tokens employ signs attested nowhere else in the Susa corpus and only once each in the full $\mathcal{S}$, suggesting that the tablet's scribe either invented new signs or employed signs specific to a peripheral scribal tradition that did not survive to significant attestation in the digitized corpus. The tablet's slot-grammar regularity ($\rho_{\text{tab}} = 0.29$, computed within this tablet alone using the normalized mutual information between sign identity and case position) is below the corpus mean of 0.38, consistent with the general PE III trend toward looser positional regularity documented in the cross-phase comparison of Proposition 10.1.

10.10 Case Study 9: An Identifier Sign in a Multi-Case Context (CDLI P009471)

Case 10.9: Recurrent Identifier Across Four Cases

Tablet P009471, a Susa PE II multi-case livestock account (7 cases, 38 sign tokens total), provides the strongest single-tablet evidence in the corpus for the H3 (identifier) sub-hypothesis of Finding 6.3. Sign PE-211 (the candidate-identifier sign examined in Case Study 5 and Worked Example 6.1) appears as the final token in 5 of the 7 cases, always immediately following a livestock numerical notation. Cases 2 and 6, the two cases without PE-211, instead close with sign PE-198 — itself a Class IV candidate-identifier sign ($\phi_{PE-198} = 0.74$, $\nu_{PE-198} = 0.69$) that appears nowhere else on this tablet but co-occurs with PE-211 on 7 other tablets in the digitized corpus. The simplest interpretation of this pattern is that P009471 records transfers between two institutional recipients (identified by PE-211 and PE-198 respectively), with the 5:2 ratio reflecting the relative sizes of their respective livestock allocations. This would make P009471 a bilateral allocation record: a document type with an obvious administrative function that requires exactly the kind of institutional-identifier sign usage posited by H3 and the candidate-identifier subset analysis of Chapter 6.

© Silvano Salvador

10.11 Case Study 10: The Largest Single Tablet (CDLI P008003)

Case 10.10: A Summary/Compilation Tablet

Tablet P008003, the largest single proto-Elamite tablet in the digitized corpus (22 administrative cases, 97 sign tokens, assigned to the summary/compilation category of Definition C.2), differs from the preceding case studies in scale and complexity. Its 97 sign tokens span 43 distinct sign types, substantially above the per-tablet average of ≈ 8 types, and 17 of the 20 Class I signs are represented. The tablet's per-token entropy, computed as the negative mean log-probability under the fitted first-order Markov model, is 3.81 bits — below the corpus mean of 4.37 bits — reflecting the higher predictability of a corpus of summary sign usage that tends toward the highest-frequency signs and the most common case-internal orderings. The case structure exhibits the clearest slot-grammar regularity in the sample ($\rho_{\text{tab}} = 0.57$, the highest of any tablet examined in this chapter), consistent with the summary/compilation document type's known function of aggregating sub-accounts into standardized format for high-level administrative review.

Note 10.5: Summary Tablets as Administrative Filters

P008003's high positional regularity ($\rho_{\text{tab}} = 0.57$) and dominance of Class I signs are consistent with an interpretation of summary/compilation tablets as administrative *filters*: documents that compress the commodity-specific detail of multiple sub-accounts (which use the full Class I–IV sign vocabulary) into a standardized summary format that retains only the highest-level commodity categories (Class I signs) in a fixed positional template. If this interpretation is correct, the summary/compilation sub-corpus provides the purest example of the slot-grammar structure central to H2: the slot positions are fixed across all cases on all summary tablets, and the sign choices within each position draw exclusively from the small Class I sub-vocabulary of universal administrative classifiers.

10.12 Case Study 11: A Peripheral-Site Discovery Tablet (CDLI P011204)

Case 10.11: A Tepe Yahya Tablet and Its Network Position

Tablet P011204 from Tepe Yahya (site code YAH, Appendix C), one of the furthest-flung find-sites in the proto-Elamite administrative network at ≈ 500 km from Susa, raises the question of how the Yahya sub-corpus relates to the Susa core. The tablet's non-numerical sign content (6 cases, 21 tokens, 11 distinct sign types) is entirely drawn from Class I and Class II signs, with not a single Class III or Class IV sign represented. This is remarkable in the context of the site-specificity finding of Proposition D.3, which established that Class IV signs are substantially more site-specific than Class I signs: if Tepe Yahya maintained an independent local scribal tradition diverging from the Susa standard, one would expect to see a higher proportion of rare, potentially site-specific signs on Yahya tablets. The absence of such signs on P011204 is consistent with the alternative hypothesis that Tepe Yahya scribes operated at a high level of institutional abstraction, recording only the broadest commodity categories in the shared pan-network vocabulary, without generating the hyper-local commodity sub-type records that would require Class III–IV signs. The Yahya sub-corpus as a whole ($n \approx 110$ tablets, Ap-

© Silvano Salvador

pendix C) shows a statistically significant under-representation of Class IV signs relative to the Susa corpus (proportion test: $\hat{p}_{IV,\text{Yahya}} = 0.08$ vs. $\hat{p}_{IV,\text{Susa}} = 0.16$; $z = -3.4$, $p < 0.001$), consistent with this high-level-summary interpretation.

10.13 Case Study 12: Reconstructing a Damaged Administrative Sequence

Case 10.12: Kneser–Ney Imputation on Tablet P007982

Tablet P007982 carries one of the most extensively damaged case sequences in the digitized corpus: of its 6 original cases (estimated from case-boundary ruling traces), 4 are partially or fully illegible, and only cases 1 and 5 are completely legible. Case 1 reads: PE-001 PE-047 PE-312 PE-001 (4 tokens); Case 5 reads: PE-001 PE-088 PE-001 (3 tokens). The Kneser–Ney imputation protocol of Section 10.6 (Proposition 10.2) and Definition 4.3 is applied to infer the most probable sign sequences for Cases 2–4 and 6, conditional on the first-order Markov model fitted to the full corpus. Under this imputation, Cases 2–4 all begin with PE-001 (posterior probability > 0.80 for this initial sign, given its role as a case-boundary marker established in Case Study 1) and end with PE-001 (posterior probability > 0.60), with differing medial signs in each case. The reconstructed tablet is consistent with a five-entry grain account (PE-001 as account-type marker, medial signs as commodity-grade or quantity qualifiers, final PE-001 as case-close marker), which would be structurally identical to Case Study 1 tablet P008832. This imputed structure is offered not as a decipherment but as an illustration of what the first-order Markov model predicts for damaged sequences, and as a demonstration that the most probable imputed structures are administratively interpretable.

Corollary 10.2: Extended Case-Study Synthesis

The twelve case studies of this chapter collectively demonstrate that the corpus-level statistical properties established in Chapters 3–7 manifest coherently at the level of individual tablets. Case Studies 1, 7, 10, and 12 illustrate the slot-grammar structure of H2; Case Studies 2 and 8 track its development across phases; Case Studies 3, 5, 9, and 11 document cross-site regularities and identifier-sign behavior; Case Study 6 validates the missing-data handling protocol; and Case Study 4 defines the limits of small-site inference. No individual tablet contradicts the corpus-level H2-favoring posterior of Corollary 7.1; several provide tablet-level corroboration that reinforces it.

© Silvano Salvador

Chapter 11

Formal Notation and Symbol Glossary

Note A.1: Scope

This appendix consolidates the mathematical notation introduced across Chapters 1–9, organized by topic area, for reference. Notation specific to a single worked example is generally not repeated here; consult the originating section for context.

11.1 Corpus and Sign-System Notation

Symbol	Meaning
$\mathcal{C}_{PE}$	The proto-Elamite corpus (Definition 1.2)
$\mathcal{N}$	Numerical sub-corpus, $\approx 14{,}000$ tokens
$\mathcal{S}$	Non-numerical sub-corpus, $\approx 10{,}000$ tokens, $\lvert\Sigma\rvert = 637$ signs
$\mathcal{R}$	The six-system reference corpus (Definition 5.1)
$\mathcal{R}_1, \ldots, \mathcal{R}_6$	Individual reference systems: proto-cuneiform, Linear Elamite, Linear B, Indus Valley, pure tallying, synthetic semasiography
Σ	The proto-Elamite non-numerical sign inventory, $\lvert\Sigma\rvert = 637$
N_{PE}	Total corpus token count, $\approx 24{,}000$
$N_{\mathcal{S}}$	Non-numerical sub-corpus token count, $\approx 10{,}000$
PE I, PE II, PE III	The three internal chronological phases (Section 2.3)

11.2 Information-Theoretic Notation

Symbol	Meaning
H_0	Zeroth-order (maximum) entropy, $\log_2 \lvert\Sigma\rvert$
H_1	First-order (unigram-frequency-weighted) Shannon entropy
H_k	kth-order conditional entropy
H_∞	Asymptotic entropy plateau as $k \to \infty$ (estimated via extrapolation)

Symbol	Meaning
H_∞ / H_0	Entropy plateau ratio, the principal restricted-register diagnostic
α_Z	Zipf-Mandelbrot exponent fitted to the sign-frequency rank distribution
ρ_{pos}	Positional regularity, $I(S;\text{Pos})/H(S)$
k^*	Effective model-order ceiling (Proposition 4.2)
$T_{Y\to X}$	Transfer entropy from process Y to process X (Definition 4.4)
ΔT	Directional transfer-entropy asymmetry, $T_{A\to B} - T_{B\to A}$
$H(B\|\hat{p}_A)$	Cross-entropy of corpus B under a model trained on corpus A
$D_{KL}(P\|Q)$	Kullback–Leibler divergence
$\mathbf{T}$	Sign-to-sign transition probability matrix (Definition 4.1)
PP_k	Perplexity of a k-gram model

11.3 Sign-Level and Positional Statistics

Symbol	Meaning
ϕ_s	Positional fixity index for sign s (Proposition 6.3)
ν_s	Numerical co-occurrence rate for sign s (Proposition 6.3)
$d(\cdot,\cdot)$	Euclidean distance in standardized classification space (Finding 5.1)

11.4 Bayesian and MDL Notation

Symbol	Meaning
H1, H1′, H2, H3, H4	The five candidate generative hypotheses (Definitions 6.1–6.4, Note 7.1)
$\pi(H_i)$	Elicited prior probability of hypothesis H_i
$\hat{\mathcal{L}}(H_i)$	Composite likelihood of hypothesis H_i (Definition 7.2)
K_{ij}	Pairwise Bayes factor, $\hat{\mathcal{L}}(H_i)/\hat{\mathcal{L}}(H_j)$
$P(H_i \mid \text{data})$	Posterior probability of hypothesis H_i
$\Delta\text{MDL}(H_i)$	Minimum Description Length differential relative to naive baseline (Definition 7.3)
$L(\text{model}_i)$	Description length of the generative model under H_i
$L(\text{data} \mid \text{model}_i)$	Negative log-likelihood of the corpus under H_i's model
$p_{KN}(s \mid c_{k-1})$	Kneser–Ney smoothed conditional sign probability (Definition 4.3)

© Silvano Salvador

Chapter 12

Statistical Methods and Computational Details

Note B.1: Scope and Reproducibility

This appendix provides estimator definitions, bias-correction procedures, and computational parameters underlying the analyses of Chapters 2–7, at a level of detail sufficient for independent reproduction from the digitized OPEP/CDLI corpus described in Note 1.1. Code and intermediate data products are organized following the same reproducibility protocol used for the companion Voynich and Linear A monographs.

12.1 Entropy Estimators and Bias Correction

All entropy estimates in this monograph use the plug-in (maximum-likelihood) estimator with Miller–Madow bias correction (Definition 3.2). For the proto-Elamite non-numerical sub-corpus at the bigram level ($k = 2$), the expected Miller–Madow correction is

$$\hat{H}_2^{\text{MM}} - \hat{H}_2 = \frac{|\Sigma_2^{(\text{obs})}| - 1}{2N_{\mathcal{S}}} \approx \frac{4{,}012 - 1}{2 \times 9{,}500} \approx 0.211 \text{ bits},$$

where $|\Sigma_2^{(\text{obs})}|$ is the number of observed distinct bigram types and $N_{\mathcal{S}} \approx 9{,}500$ is the number of non-final bigram positions in the non-numerical sub-corpus. This correction is substantially larger, in relative terms, than the corresponding correction for the Linear A corpus (0.091 bits at $N_{LA} = 7{,}362$), reflecting the larger proto-Elamite sign inventory ($|\Sigma| = 637$ versus Linear A's considerably smaller sign-group inventory) and the correspondingly sparser bigram table. The Chao–Shen entropy estimator (Chao and Shen, 2003) is used throughout as a robustness check; it produces estimates within 0.07 bits of the Miller–Madow corrected values for all entropy figures reported in Chapters 3–5.

12.2 Bootstrap Confidence Intervals

All confidence intervals in this monograph are bootstrap CIs based on $B = 10{,}000$ resamples unless otherwise stated. The bootstrap resamples administrative *cases* (Definition 1.2) with replacement, preserving the within-case sign sequence and case-boundary structure; this is the appropriate resampling unit because cases, not individual sign tokens, constitute the natural unit of statistical independence in an administrative corpus where successive cases on a

tablet may record related but distinct transactions. The classification-space ellipsoid of Proposition 5.1 and the transfer-entropy confidence intervals of Findings 4.2 and the corresponding Note 4.1 are computed via this case-level bootstrap. The 95% bootstrap CI throughout uses the bias-corrected and accelerated (BCa) method of Efron and Tibshirani (1993), which corrects for both bias and skewness in the bootstrap sampling distribution; the bias-correction constant $\hat{z}_0$ and acceleration constant $\hat{a}$ are estimated separately for each reported statistic via the standard jackknife-after-bootstrap procedure.

12.3 Permutation Tests

All permutation p-values reported in this monograph are based on $M = 100{,}000$ permutation replicates. For the cross-entropy and sign-similarity tests of Chapters 4 and 8 (Finding 4.1, Worked Example 8.1), the permutation procedure randomly reassigns sign-frequency labels while preserving the total token count and the marginal frequency distribution of the comparison corpus, in order to construct a null distribution appropriate to testing divergence between two empirically observed frequency distributions rather than between a fitted model and data. For the bimodality test underlying Finding 6.3 (Hartigan's dip test), critical values are obtained via the standard dip-test bootstrap calibration (Hartigan and Hartigan, 1985) rather than asymptotic approximation, given the moderate sample size ($n = 637$ signs).

12.4 Corpus-Size Matching Experiments

To distinguish genuine structural differences from artifacts of differing corpus size in the cross-system comparisons of Chapter 5 (Finding 5.1), all six reference systems are subsampled to a matched size of $N^* = 10{,}000$ tokens prior to computing $|\Sigma|$, H_∞ / H_0, and ρ_{pos}, with 200 independent subsamples drawn per system and the reported statistics taken as subsample means. For proto-cuneiform ($\mathcal{R}_1$, full corpus $N \approx 137{,}000$) and Linear B ($\mathcal{R}_3$, full corpus $N \approx 60{,}000$), this matching is a genuine downsampling; for Linear Elamite ($\mathcal{R}_2$, full corpus $N \approx 1{,}600$), the matched-size comparison is instead achieved by upward extrapolation of the entropy plateau using the same bias-correction framework applied to the small-sample proto-Elamite estimates themselves, flagged in Table 5.1's accompanying notes (not reproduced in the main text) as carrying wider uncertainty than the directly-subsampled systems.

12.5 The Bayesian Chronological Model

The Bayesian chronological model underlying Chapter 2's three-phase periodization treats each tablet's absolute date τ_i as drawn from a posterior

$$p(\tau_i \mid \text{strat}_i, \text{glyptic}_i) \propto p(\text{strat}_i \mid \tau_i)\, p(\text{glyptic}_i \mid \tau_i)\, p(\tau_i),$$

with the stratigraphic likelihood $p(\text{strat}_i \mid \tau_i)$ modeled as a layer-specific uniform distribution over the excavator's published date range for each stratigraphic unit, and the glyptic likelihood $p(\text{glyptic}_i \mid \tau_i)$ modeled as a Gaussian centered on the seriated date of the associated seal style, with standard deviation set from the published seriation's internal consistency statistics. The prior $p(\tau_i)$ is a weakly informative uniform distribution over the broad excavation-attested date range (3300–2800 BCE). Posterior sampling is performed via Hamiltonian Monte Carlo (No-U-Turn sampler), with convergence assessed via $\hat{R} < 1.01$ and effective sample size $> 1{,}000$ for all reported posterior date estimates, following the diagnostic standards of Vehtari et al. (2021).

© Silvano Salvador

12.6 Transfer Entropy Estimation Details

The transfer entropy computations of Section 4.4 (Findings 4.2, Note 4.1) require estimating four-way joint distributions $p(x_{t+1}, x_t, y_t)$ over commodity-class sign choices at matched tablet positions across site or phase pairs. Given the modest sample sizes available per commodity class ($n \in [50, 340]$ tablets per site/phase/commodity combination), all transfer-entropy estimates use the James–Stein shrinkage estimator for the underlying joint frequency table (shrinking toward the independence-structured table) rather than the raw plug-in estimator, in order to control variance at the cost of a small, analytically bounded bias; the bias bound is incorporated into the bootstrap confidence intervals reported for ΔT throughout Section 4.4.

12.7 Kneser–Ney Discount Parameter Estimation

The Kneser–Ney discount parameter d used in the higher-order smoothed probability estimates of Definition 4.3 (required for the Bayesian synthesis of Chapter 7, which depends on probability estimates beyond the $k^* = 1$ ceiling established for unsmoothed models) is estimated by leave-one-case-out cross-validation over the non-numerical sub-corpus. The leave-one-out log-likelihood

$$\ell_{\text{LOO}}(d) = \sum_{(s,s') \in \mathcal{D}} \log p_{KN}(s' \mid s; d)$$

is maximized at $\hat{d} = 0.71$ (95% CI $[0.64, 0.78]$ by profile likelihood), close to the canonical default of 0.75 and to the corresponding Linear A estimate of 0.74, suggesting that the optimal discount parameter is relatively stable across undeciphered-script corpora of comparable size and is not itself strongly diagnostic for hypothesis discrimination.

12.8 Classification-Space Distance and the Covariance Estimate

The Euclidean distances reported in Finding 5.1 are computed in a standardized (unit-variance, zero-mean) version of the three-statistic space $(|\Sigma|, H_\infty/H_0, \rho_{pos})$, with standardization constants estimated from the pooled distribution across all seven systems (six reference systems plus proto-Elamite). A Mahalanobis-distance robustness check, using the sample covariance matrix of the three-statistic vectors across the seven systems with Ledoit–Wolf shrinkage regularization (appropriate given the small number of systems, $n = 7$, relative to the dimension, $d = 3$), reproduces the same qualitative ordering of reference-system distances as the simple Euclidean computation, with the proto-cuneiform and synthetic-semasiography systems remaining the two closest comparanda under both metrics.

12.9 The MDL Computation in Detail

The model description lengths $L(\text{model}_i)$ reported in Finding 7.2 are computed using a two-part MDL scheme (Rissanen, 1996): each model parameter is charged $\frac{1}{2}\log_2 N_S$ bits (the standard asymptotic per-parameter coding cost under a normal approximation to the parameter's sampling distribution), multiplied by the number of free parameters in each hypothesis's generative model — $637^2 - 1$ for the unrestricted bigram model of H1, $|\Sigma_{\text{slot}}| \times |\text{slots}| - 1$ for the slot-grammar model of H2 (with $|\Sigma_{\text{slot}}| \approx 120$ effective per-slot sub-vocabularies estimated from the clustering analysis underlying Section 5.1's synthetic semasiography control), and intermediate parameter counts for H1′, H3, and H4 as detailed in the project's computational appendix. The data-encoding term $L(\text{data} \mid \text{model}_i)$ is the negative log-likelihood of the

© Silvano Salvador

observed corpus under each fitted model, computed using the Kneser–Ney smoothed probabilities of Section B.6 for all models requiring probability estimates beyond directly observed frequencies.

12.10 Software and Computational Environment

All statistical computations were performed in Python 3.11 using NumPy, SciPy, and a custom information-theoretic toolkit shared with the companion Voynich and Linear A analyses (ensuring identical entropy, Zipf-fitting, and MDL estimator implementations across all three undeciphered-script monographs, so that cross-script comparisons reported in this and the companion volumes are not confounded by implementation-specific estimator differences). Bayesian posterior sampling for the chronological model of Section B.5 was performed using a Hamiltonian Monte Carlo sampler with default NUTS tuning parameters (target acceptance rate 0.8, 4 chains of $2{,}000$ post-warmup iterations each).

12.11 Convergence Diagnostics for the Bayesian Chronological Model

Reproducibility of the Bayesian chronological posteriors reported throughout Chapter 2 depends on adequate MCMC convergence, which this section documents explicitly rather than merely asserting.

- **Gelman–Rubin statistic:** $\hat{R} < 1.01$ for all model parameters across the four independent chains, well within the conventional 1.01 convergence threshold.
- **Effective sample size:** minimum ESS across parameters is $1{,}840$ (of $8{,}000$ total post-warmup draws), with the stratigraphy-layer hyperparameters showing the lowest ESS due to their comparatively diffuse posterior, still well above the conventional minimum-adequacy threshold of 400.
- **Divergent transitions:** 3 divergences out of $8{,}000$ total post-warmup iterations (0.04%), all occurring in the same chain during exploration of the low-density tail of the PE I/PE II boundary region; re-running with an increased target acceptance rate of 0.95 eliminates all divergences without materially changing any posterior summary, confirming the divergences reflect a minor sampling inefficiency rather than a model misspecification.

12.12 Estimator Validation Against Known-Ground-Truth Synthetic Corpora

Several estimators used in this monograph — the Miller–Madow-corrected entropy estimator (Definition E.1), the Zipf–Mandelbrot fitting routine, and the CTW compressor of Definition 4.6 — are validated not only against the proto-Elamite corpus itself but against synthetic corpora of known ground-truth generative structure, a validation step performed once and applied uniformly across all three companion monographs.

Note B.7: The Synthetic Validation Suite

For each estimator, 500 synthetic corpora are generated from a known generative process (a fixed-order Markov chain of specified order and known true entropy, for the entropy and CTW estimators; a known Zipf–Mandelbrot distribution with specified parameters, for the Zipf fitting routine) at a range of corpus sizes spanning the proto-Elamite corpus's own scale ($N \in \{2{,}000, 5{,}000, 10{,}000, 20{,}000\}$ tokens). The estima-

© Silvano Salvador

tor's bias and root-mean-squared error against the known ground truth are reported at each corpus size, providing the empirical basis for the bias-correction terms and confidence-interval widths used throughout this monograph rather than relying solely on asymptotic theory, which is known to be a poor guide at the corpus sizes available for proto-Elamite (Proposition 1.1).

At the proto-Elamite corpus's own approximate scale ($N \approx 10{,}000$ tokens for the non-numerical sub-corpus), the Miller–Madow entropy estimator shows residual bias below 0.03 bits and RMSE below 0.09 bits against known ground truth; the Zipf–Mandelbrot exponent estimator shows bias below 0.02 and RMSE below 0.06; and the CTW compression-rate estimator shows bias below 0.04 bits/sign. These figures are the basis for the uncertainty terms attached to the corresponding point estimates throughout Chapters 2–4.

© Silvano Salvador

Chapter 13

Corpus Catalogue: Sites, Phases, and Document Categories

Note C.1: Purpose of This Appendix

This appendix provides a summary catalogue of the proto-Elamite corpus underlying the analyses of Chapters 2–7 and 10, organized by find-site, internal chronological phase, and administrative document category. It is intended as a reference inventory rather than a full archaeological publication of the corpus; readers seeking complete excavation reports and individual tablet photographs should consult the CDLI and OPEP digital archives cited in Note 1.1.

13.1 Catalogue by Find-Site

Definition C.1: Site Codes Used Throughout This Appendix

Four site codes are used in the tabulations below: **SUS** (Susa, the type-site and largest single find-spot); **ANS** (Tal-i Malyan, ancient Anshan); **SIA** (Tepe Sialk); **YAH** (Tepe Yahya); and **OTH** (a residual category of smaller find-spots, individually contributing fewer than ten tablets each to the digitized corpus).

Site	Tablets	Non-num. tokens	% of $\mathcal{S}$	Notes
SUS (Susa)	$\approx 1{,}050$	$\approx 6{,}400$	64%	Type-site; PE I–III all attested; primary scribal-innovation locus per Finding 4.2
ANS (Anshan)	≈ 280	$\approx 1{,}750$	17.5%	PE II–III dominant; strong sign-inventory overlap with Susa (Case Study 3)
SIA (Sialk)	≈ 95	≈ 610	6.1%	Smaller corpus; statistical power limitations noted in Note 10.3
YAH (Tepe Yahya)	≈ 110	≈ 690	6.9%	Source of the Damerow–Englund (1989) foundational publication

Site	Tablets	Non-num. tokens	% of $\mathcal{S}$	Notes
OTH (smaller sites)	≈ 65	≈ 550	5.5%	Individually below the threshold for site-specific statistical treatment
Total	$\approx 1{,}600$	$\approx 10{,}000$	100%	

13.2 Catalogue by Internal Phase

Phase	Tablets (est.)	Approx. date range	Characteristic features
PE I	≈ 280	3200–3100 BCE	Looser case-length structure (Case Study 2); PE-001 framing pattern not yet stabilized; earliest attested layer at Susa
PE II	≈ 980	3100–3000 BCE	Peak corpus density; most regular slot-grammar structure (Case Studies 1, 5); primary basis for the corpus-wide statistics of Chapters 3–5
PE III	≈ 340	3000–2900 BCE	Sign-repertoire drift and site-level fragmentation (Finding 2.1, Note 4.1); weakened cross-phase transfer entropy

Note C.2: Phase Assignment Uncertainty

Phase assignment for approximately 12% of digitized tablets rests on stratigraphic context alone, without independently dateable glyptic material (Section 2.2); these tablets are assigned to the phase with highest posterior probability under the Bayesian chronological model of Appendix B.5 but carry correspondingly wider date uncertainty. All phase-stratified statistics reported in Chapters 2 and 4 were verified to be robust to exclusion of this uncertain-phase subset (a robustness check not separately tabulated, following the same protocol as the missing-data sensitivity analysis of Proposition 10.2).

13.3 Catalogue by Administrative Document Category

Definition C.2: Document Categories

Following the functional classification scheme established in the foundational proto-Elamite epigraphic literature (Englund, 1996, 2004) and adopted by the OPEP digital corpus, tablets are classified into five non-exclusive functional categories on the basis of the numerical notation system employed and the accompanying non-numerical sign pattern:

- **Livestock accounts:** records employing the sexagesimal livestock counting system, frequently co-occurring with the candidate-identifier signs of Finding 6.3 (Case Study 5);

© Silvano Salvador

- **Grain/capacity accounts:** records employing capacity system S, the largest single category by tablet count, exemplified by Case Study 1;
- **Discrete-object accounts:** records employing capacity system N for counting discrete manufactured or processed goods;
- **Labor/personnel records:** a smaller category, generally exhibiting the highest concentration of candidate-identifier signs per tablet;
- **Summary/compilation tablets:** a small category of larger tablets aggregating multiple sub-accounts, often the source of the longest individual administrative cases in the digitized corpus.

Category	Tablets (est.)	% of corpus	Representative case study
Livestock accounts	≈ 420	26%	Case Study 5 (P009154)
Grain/capacity accounts	≈ 610	38%	Case Study 1 (P008832)
Discrete-object accounts	≈ 290	18%	Case Study 3 (P010447)
Labor/personnel records	≈ 165	10%	—
Summary/compilation tablets	≈ 115	7%	—

Proposition C.1: Category Is Not a Strong Predictor of the Restricted-Register Anomaly

A category-stratified re-estimation of the entropy plateau ratio H_∞/H_0 (the central restricted-register diagnostic of Section 3.3) finds only modest variation across the five document categories, ranging from 0.62 (labor/personnel records) to 0.69 (summary/compilation tablets), with overlapping bootstrap confidence intervals in all pairwise comparisons. This near-uniformity across functionally distinct document categories is itself informative: it indicates that the restricted-register anomaly is a property of the non-numerical sign system as a whole rather than an artifact specific to any one administrative sub-domain, somewhat strengthening the corpus-general H2 interpretation favored by the Bayesian synthesis of Chapter 7 over an alternative account in which different document categories might encode language with differing degrees of restriction.

13.4 Digitization Status and Data Provenance

Note C.3: Digitization Status

Of the catalogued $\approx 1{,}600$ tablets and fragments, $\approx 80\%$ are fully digitized and transliterated in the OPEP corpus used throughout this monograph (Note 1.1); a further $\approx 12\%$ are partially digitized (transliterated but not yet cross-checked against high-resolution imaging); and the remaining $\approx 8\%$ are catalogued by find-context only, pending future transliteration. All statistics reported in this monograph are computed from the fully and partially digitized subsets ($\approx 92\%$ of the catalogued corpus), with the missing-data handling protocol of Section 10.6 and Proposition 10.2 applied uniformly to partially legible material within this subset.

© Silvano Salvador

Chapter 14

Sign-Level Distributional Profiles: A Complete Inventory Analysis

14.1 Motivation and Scope of the Sign-Level Analysis

The corpus-level statistics of Chapters 3–5 treat the proto-Elamite non-numerical sign sequence as a single statistical object, pooling across all sign types to produce aggregate entropy, Zipf, and positional regularity measures. This appendix descends to the sign level, providing a systematic distributional analysis of individual sign types and sign classes. The motivation is threefold. First, the hypothesis evaluation of Chapter 6 predicts qualitatively different *within-inventory* distributional patterns under H2 versus H1′: a constrained semasiography (H2) should exhibit a small set of ubiquitous high-frequency signs covering all sites and phases, a large set of mid-frequency commodity-specific signs, and a long tail of rare signs that may be site-specific or scribe-specific; a restricted natural language (H1′) should instead exhibit a bipartite structure closer to the closed-class/open-class distinction of natural lexicons, with grammatical markers forming one cluster and content items another. Second, the 73-sign candidate-identifier subset isolated in Finding 6.3 was identified via corpus-level indicators (ϕ_s and ν_s); this chapter examines the remaining 564 non-identifier signs more finely, asking whether they partition into interpretable sub-classes. Third, the cross-site and cross-phase distributional comparisons of Chapters 2 and 4 were conducted at the corpus level; here they are repeated sign by sign, producing a matrix of per-sign Jensen–Shannon divergences across sites that directly operationalizes the shared-versus-local-vocabulary question central to the H4 evaluation.

Definition D.1: Sign-Frequency Classes

The 637-sign reduced inventory Σ^*_{PE} is partitioned into four frequency classes based on corpus token rank:

- **Class I (ubiquitous signs):** rank 1–20; account for $\approx 38\%$ of all non-numerical tokens; attested at all four principal sites.
- **Class II (common signs):** rank 21–100; account for $\approx 31\%$ of tokens; attested at ≥ 3 principal sites.
- **Class III (uncommon signs):** rank 101–300; account for $\approx 20\%$ of tokens; site distribution heterogeneous.
- **Class IV (rare signs):** rank 301–637; account for $\approx 11\%$ of tokens; many attested at only one site or one phase.

14.2 Class I: The Ubiquitous Sign Layer

Proposition D.1: Class I Signs as Structural Anchors

The 20 Class I signs share three statistical properties that jointly distinguish them from all lower-ranked signs: (i) their positional fixity index ϕ_s is uniformly below 0.50 (mean $\phi = 0.31$, standard deviation 0.08), indicating that they are *not* position-fixed and therefore do not function as slot-boundary markers in the sense of the candidate-identifier subset; (ii) their cross-site Jensen–Shannon divergence is uniformly low, with $\mathrm{JSD}(\text{Susa} \| \text{Anshan}) < 0.05$ bits for all 20 signs; and (iii) their rank order is stable across the three chronological phases (Kendall $\tau = 0.87$ between PE I and PE II rank orders, $\tau = 0.82$ between PE II and PE III), indicating that the core high-frequency layer is both geographically and temporally stable.

Under the H2 (constrained semasiography) interpretation, this stability profile is the expected signature of a small set of universal administrative classifiers — signs that categorize broad commodity domains ("grain-class", "livestock-class", "labor-class") in a way that is meaningful across all sites and phases precisely because it indexes stable categories of the proto-Elamite economic system rather than locally variable spoken-language categories. The stability observed for Class I signs provides direct per-sign corroboration of the cross-site transfer-entropy result of Finding 4.2 and the cross-phase standardization finding of Chapter 2.

Worked Example D.1: Signs PE-001, PE-002, and PE-003

The three highest-frequency signs in $\mathcal{S}$ — PE-001 (4.8% of tokens, rank 1), PE-002 (3.1%, rank 2), and PE-003 (2.7%, rank 3) — together account for 10.6% of all non-numerical tokens. All three exhibit the cross-site and cross-phase stability of Proposition D.1. Their mutual co-occurrence pattern is asymmetric: PE-001 and PE-002 co-occur within the same administrative case at a rate $3.4\times$ above the independence baseline ($\chi^2 = 218.4$, $df = 1$, $p < 0.001$), while PE-001 and PE-003 co-occur at a rate only $1.2\times$ above baseline ($p = 0.08$, not significant). This asymmetry suggests that PE-001 and PE-002 index the same broad commodity domain while PE-003 indexes a distinct one, consistent with an interpretation of PE-001/PE-002 as grain-class classifiers and PE-003 as a livestock-class classifier — though the corpus-internal evidence alone cannot confirm this semantic reading.

14.3 Class II: The Common-Sign Layer and Commodity Sub-Vocabularies

The 80 Class II signs (ranks 21–100) exhibit a substantially more heterogeneous distributional profile than Class I. A hierarchical clustering of their site-specific frequency vectors (Ward linkage, Euclidean distance in normalized frequency space) recovers four clusters of comparable size ($n \approx 18$–22 signs per cluster), corresponding roughly to the four principal administrative document categories of Appendix C: a livestock-account cluster, a grain/capacity cluster, a discrete-object cluster, and a labor/personnel cluster. This cluster structure is not imposed by the document-category classification itself (which uses numerical notation to assign categories, not non-numerical sign content) and therefore constitutes an independent piece of evidence for the semantic coherence of the Class II sign layer.

© Silvano Salvador

Proposition D.2: Commodity-Domain Clustering of Class II Signs

A silhouette analysis of the four-cluster solution for Class II signs (mean silhouette coefficient $\bar{s} = 0.61$) confirms that the four-cluster partition is substantially better than either a two-cluster or six-cluster solution ($\bar{s} = 0.44$ and 0.51 respectively). The four recovered clusters correlate significantly with the four document categories of Appendix C: Cramér's $V = 0.58$ between cluster membership and document-category majority affiliation ($\chi^2 = 41.2$, $df = 9$, $p < 0.001$). This correlation, obtained by comparing a sign-frequency-based clustering against an independently-derived document classification, constitutes the strongest purely corpus-internal evidence available in this monograph that the non-numerical sign system carries commodity-domain information.

14.4 Class III and IV: Rare Signs, Local Vocabulary, and the Long Tail

Note D.1: The Interpretation of Rare Signs

The 337 Class III and IV signs collectively account for only 31% of non-numerical tokens despite comprising 53% of the sign inventory. Two competing interpretations apply to rare signs: (a) under H2, they represent a fine-grained commodity vocabulary covering specific sub-types, quantities, or institutional contexts too rare to receive a high frequency in the surviving corpus; (b) under H1′, they represent the open class of a natural lexicon, with natural laws of word frequency (Zipf's law for the body of the distribution plus Heap's law for vocabulary growth) determining that a substantial fraction of vocabulary items must be rare. These interpretations are not distinguishable by sign-level frequency statistics alone, since Zipf's law holds for both natural-language vocabularies and for combinatorially rich semasiographic systems. The distinguishing evidence is instead cross-site: if rare signs are open-class vocabulary items, their distribution across sites should be roughly uniform (all sites have access to the same lexicon); if they are hyper-local commodity sub-types, they should be substantially more site-specific than Class I signs.

Proposition D.3: Rare Signs Are Substantially More Site-Specific

For the 200 lowest-frequency signs in Σ^*_{PE} (Class IV plus the bottom half of Class III), the mean cross-site Jensen–Shannon divergence between Susa and Anshan is $\text{JSD}_{IV} = 0.38$ bits — more than seven times larger than the corresponding value for Class I signs (0.05 bits, Proposition D.1) and significantly above the Class II value of 0.12 bits (t-test, $p < 0.001$). Of these 200 signs, 87 (43.5%) are attested at only one of the four principal sites, compared to 0% of Class I signs. This strong site-specificity of the long tail supports the H2 commodity-sub-vocabulary interpretation over the H1′ open-class-lexicon interpretation, since the latter predicts a common lexical inventory available at all sites, not a fragmented local vocabulary concentrated in the rare-sign tier.

© Silvano Salvador

14.5 The Sign Co-Occurrence Graph and Its Structural Properties

Definition D.2: The Sign Co-Occurrence Graph G_{PE}

The sign co-occurrence graph $G_{PE} = (V, E, w)$ is defined with vertex set $V = \Sigma^*_{PE}$ (637 nodes), edge set E connecting pairs of signs that co-occur within the same administrative case at least twice, and edge weight $w_{ij} = \text{PMI}(s_i, s_j)$ (pointwise mutual information) measuring the strength of the co-occurrence association beyond chance expectation:

$$\text{PMI}(s_i, s_j) = \log_2 \frac{p(s_i, s_j)}{p(s_i)\, p(s_j)},$$

where $p(s_i, s_j)$ is the probability that signs s_i and s_j co-occur in the same case, and $p(s_i)$, $p(s_j)$ are their marginal token probabilities. Edges with PMI < 0 (co-occurrence below chance expectation) are excluded.

Proposition D.4: The Co-Occurrence Graph Has Community Structure

Applying the Louvain community detection algorithm to G_{PE} (with edge weights w_{ij}, resolution parameter $\gamma = 1.0$), the algorithm recovers $Q = 7$ communities with a modularity score of $\hat{Q} = 0.41$, substantially above the null-model expectation of $Q^* \approx 0.05$ for a random graph of matching degree distribution (permutation test, $p < 0.001$). The seven communities range in size from 41 to 128 signs, with the two largest communities closely corresponding to the livestock-cluster and grain-cluster of Class II signs (Proposition D.2), and the remaining five communities capturing more specialized or ambiguous sign groupings. Cross-community edges are predominantly between adjacent large communities, consistent with the administrative reality that some tablet types (summary accounts) record multiple commodity domains simultaneously.

The community structure of G_{PE} provides a sign-level visualization of the same domain-clustering result established at the aggregate level by Proposition D.2. More importantly, it constrains the form of the slot-grammar model central to H2 (Definition 6.2): the seven communities correspond approximately to the commodity-domain sub-vocabularies from which each case-position slot draws its sign, suggesting a richer grammar than the simplified three-slot template of Definition 4.5 but one that remains well within the semasiographic rather than linguistic parameter range.

Worked Example D.2: Community 3 as a Candidate Textile Sub-Vocabulary

Community 3 ($n = 67$ signs) contains 11 of the 19 non-numerical signs attested on the Case Study 3 Anshan textile tablet (P010447) and 8 of 12 signs on a second Anshan textile tablet (CDLI P010512, not otherwise discussed in this monograph). The community's mean positional regularity within cases ($\rho_{\text{comm},3} = 0.44$, above the corpus-wide $\rho_{pos} = 0.38$) and its concentration in discrete-object document types (Appendix C, 61% of Community 3 tokens appear on discrete-object tablets) are consistent with a textile/manufactured-goods sub-vocabulary: fixed-position production-type qualifiers, fabric-grade designators, and institutional-source markers.

© Silvano Salvador

Chapter 15

Computational Reproducibility: Code Architecture and Validation

15.1 Overview of the Computational Pipeline

All quantitative results in this monograph were produced by a unified computational pipeline shared, with parameter adaptations specific to each corpus, across the three companion undeciphered-script monographs (Voynich, Linear A, and proto-Elamite). This chapter documents the architecture of that pipeline in sufficient detail to allow independent reproduction from the publicly available OPEP/CDLI digital corpus described in Note 1.1, complementing the estimator-level specification of Appendix B with a description of the data-flow and code-organization decisions that determine how the estimators are applied.

Definition E.1: The Computational Pipeline Stages

The pipeline is organized into six sequential stages:

1. **Corpus ingestion**: CDLI ATF-format transcriptions are parsed into a structured corpus object encoding the case hierarchy of Definition 1.2.
2. **Sign normalization**: OPEP sign identifiers (PE-nnn) are mapped to a normalized integer vocabulary after exclusion of damaged-position markers and numerical notation tokens.
3. **Sub-corpus partition**: tokens are partitioned into $\mathcal{N}$ (numerical) and $\mathcal{S}$ (non-numerical) using the OPEP numerical-sign flag, and further partitioned by site and phase using the Bayesian phase assignments of Appendix B.5.
4. **Statistical estimation**: entropy, Zipf, n-gram, transfer-entropy, and classification-space statistics are computed from $\mathcal{S}$ and its sub-corpora using the estimators of Appendix B.
5. **Bootstrap and permutation resampling**: confidence intervals and p-values are computed via the resampling protocols of Appendix B.2–B.3.
6. **Bayesian synthesis**: Bayes factors and MDL differentials are computed from the output of Stage 4, using the model-description-length framework of Appendix B.9.

15.2 Sign Normalization and the Hapax Problem

Note E.1: The Challenge of Hapax Legomena

The 363 hapax legomena (signs attested exactly once in $\mathcal{S}$, excluded from Σ^*_{PE} per Definition 1.3) pose a non-trivial normalization challenge: they cannot simply be discarded, since their existence influences vocabulary-size estimates, entropy bias corrections, and MDL parameter counts. The pipeline handles hapax legomena via a dedicated "UNK" (unknown) pseudo-type added to Σ^*_{PE} with an aggregated frequency equal to the total hapax token count ($n = 363$ tokens after exclusion of the 363 *types*). All statistics reported for $|\Sigma^*_{PE}|$ in this monograph use the 637-type reduced inventory, while statistics that depend on vocabulary completeness (particularly the Miller–Madow bias correction of Appendix B.1 and the MDL parameter counting of Appendix B.9) use the 637+UNK = 638-type extended inventory.

Proposition E.1: The UNK-Handling Protocol Does Not Materially Affect Main Results

A sensitivity analysis comparing the 637-type and 638-type (UNK-inclusive) inventory treatments for the six principal diagnostic statistics of Note 6.1 finds differences of: $\Delta H_0 = +0.01$ bits (within bootstrap CI), $\Delta\alpha_Z = -0.02$ (not significant), $\Delta\rho_{pos} = 0.00$ (unchanged by construction), $\Delta k^* = 0$ (unchanged), $\Delta H(\text{PE}\|\hat{p}_{PC}) = +0.08$ bits (negligible), and $\Delta T_{\text{Susa}\to\text{Anshan}} = -0.001$ bits (negligible). The posterior ranking of Corollary 7.1 is invariant to the UNK-handling choice; the 637-type convention is adopted throughout for notational simplicity.

15.3 Validation Against Independent Published Results

Note E.2: Available External Benchmarks

Three published quantitative results on the proto-Elamite corpus provide external benchmarks against which the present pipeline can be validated: (i) Friberg's (1978) count of the capacity system N token frequency at Susa; (ii) Englund's (2004) published sign-frequency rank for the 50 most common non-numerical signs; and (iii) the OPEP project's own published inter-rater agreement statistics for sign identification (approximately 85% agreement for Class I signs, declining to 62% for Class IV signs).

Worked Example E.1: Reproducing Englund's Top-50 Frequency Rank

Applying the pipeline's Stage 2 sign normalization and Stage 3 sub-corpus partition to the full digitized OPEP corpus reproduces Englund's (2004) top-50 frequency rank with a Kendall $\tau = 0.93$ — high agreement that validates the corpus ingestion and sign normalization stages (Steps 1–2 of Definition E.1). The seven discrepancies between Englund's published rank and the pipeline's rank are attributable to three sources: (a) corpus growth since Englund's 2004 tabulation (OPEP has added approximately 180 newly digitized tablets since then, shifting ranks for mid-frequency signs); (b) one sign split by OPEP's later sign-list revision (PE-087 in Englund's numbering became PE-087a and PE-087b in the current OPEP list, distributing the token count between two entries); and (c) the pipeline's explicit exclusion of damaged-position tokens (roughly

© Silvano Salvador

4% of the corpus), which Englund's tabulation partially included.

15.4 Numerical Sub-System Validation as a Ground-Truth Check

Worked Example E.2: Entropy of the Capacity System N

As a ground-truth validation of the pipeline's entropy estimation stage (Step 4 of Definition E.1), the zeroth-order Shannon entropy of the capacity system N notations (a fully understood positional-value metrological system) is estimated: $\hat{H}_0^N = 4.18 \pm 0.07$ bits. This value matches the theoretical entropy of a positional notation system with an effective digit inventory of $2^{4.18} \approx 18$ equiprobable digits, consistent with the capacity system N's known structure (a system using approximately 15–20 graphically distinct notation signs with approximately uniform administrative frequency in the Susa grain-accounting corpus). The agreement between the empirical estimate and the theoretically-predicted range (4.0–4.4 bits based on published sign-list counts) validates the entropy estimator's bias correction at the specific corpus size of $\mathcal{N}$ ($N \approx 5{,}000$ tokens), directly analogous to the cross-validation provided by Finding 3.4 for the full dual sub-system.

15.5 Sensitivity to ATF Transcription Conventions

Note E.3: ATF Format and Transcription Uncertainty

The CDLI ATF (ASCII Transliteration Format) transcription convention introduces a systematic choice with statistical implications: when a sign is partially damaged but identifiable by its remaining strokes, it is transcribed with a certainty flag (either an explicit sign identifier or a bracketed restoration); when it is unidentifiable, it is marked with a generic damaged-position symbol. The pipeline treats all bracketed restorations as equivalent to unbracketed transcriptions (following the OPEP project's explicit editorial policy that restorations are made only when certainty is high), while all generic damage markers are excluded from statistical analysis per the protocol of Section 10.6. A sensitivity analysis using a stricter policy (excluding bracketed restorations as well) reduces N_S from 10,142 to 9,408 tokens but does not shift any headline statistic outside its already-reported bootstrap CI.

© Silvano Salvador

Chapter 16

The Archaeological and Cultural Context

16.1 The Urban Revolution and the Origins of Writing

Proto-Elamite does not emerge in a historical vacuum. It arises within — and arguably as a constitutive part of — a moment of transformative social change that archaeologists have termed the Urban Revolution: the rapid crystallization, in the late fourth millennium BCE, of large, densely populated, economically complex settlements that require new technologies of social coordination. The parallel emergence of proto-cuneiform at Uruk-period Mesopotamia and proto-Elamite at Susa-period Iran are the two most conspicuous examples of this coordination-driven writing invention, and their near-simultaneity is itself a datum requiring explanation. The present monograph's statistical analysis is explicitly agnostic on the causal question of whether the two scripts share an idea ("writing as a concept") diffused from Mesopotamia to Iran, or whether they represent structurally convergent responses to a shared administrative pressure; but the cultural context frames the statistical findings and determines what kinds of explanation are available.

Definition F.1: The Susiana Administrative Complex

The term *Susiana administrative complex* refers to the cluster of proto-Elamite find-sites on the Susiana plain and adjacent highland zones (Susa, Anshan, Sialk, Yahya) considered as a single, loosely coordinated economic network during the PE I–II period. The hypothesis that this network was administered, at least partially, through the proto-Elamite tablet system is supported by: (i) the geographic distribution of the corpus across sites at distances of 100–500 km from Susa; (ii) the shared sign conventions of the Class I sign layer (Proposition D.1); and (iii) the Susa-outward transfer-entropy asymmetry of Finding 4.2. The term "complex" is used without prejudice to the political question of whether a centralized polity administered the network or whether it operated through voluntary inter-site coordination of the kind formalized in the Network Horizon framework (Note 8.4).

16.2 Susa: Stratigraphy, Excavation History, and Source Criticism

Note F.1: The Excavation Record and Its Biases

The proto-Elamite corpus is dominated by finds from Susa, the type-site and the locus of the most extensive controlled excavations. Jacques de Morgan's expeditions (1897–1913) and the subsequent French Délégation Archéologique en Iran (DAFI) work produced the bulk of the stratigraphically contextualized material. However, this excavation history introduces a systematic source bias: the DAFI excavations prioritized contexts adjacent to the known administrative areas of Susa Levels XIV–XVII, making the Susa corpus disproportionately weighted toward a particular administrative function (palatial or temple-based grain and livestock accounting) and a particular phase (PE II, the period of maximum administrative activity in those areas). Peripheral sites (Sialk, Yahya) were excavated on a smaller scale and with less stratigraphic control. These biases do not invalidate the statistical analyses of the present monograph, since the sensitivity analyses of Chapters 2, 7, and 10 establish robustness to site-imbalanced sampling, but they must be kept in mind when interpreting any finding that compares Susa against non-Susa sites.

16.3 Seal Impressions and the Glyptic Record

Cylinder seals and their clay impressions constitute the most important dating tool available for the proto-Elamite corpus (Sections 2.2–2.3) and also an independent channel of administrative information parallel to the tablet sign sequence. Understanding the relationship between glyptic (seal-impression) data and the non-numerical sign sequences examined in this monograph is therefore both practically important (for the chronological model of Chapter 2) and theoretically significant (for the hypothesis-evaluation framework of Chapter 6).

Proposition F.1: Glyptic and Sign-Sequence Data Are Partially Independent

A test of statistical independence between the glyptic style class $\sigma \in \{1, \ldots, 7\}$ (Definition 2.2) and the Class I sign distribution on co-associated tablets (the 241 tablets carrying both a legible cylinder-seal impression and ≥ 5 non-numerical sign tokens) finds a modest but statistically significant association: Cramér's $V = 0.14$ ($\chi^2 = 28.4$, $df = 6$, $p < 0.001$). The association is entirely consistent with the shared temporal information (both glyptic style and sign frequency drift across phases), and the residual association after controlling for phase ($V_{\text{partial}} = 0.07$, $p = 0.08$, not significant) suggests that glyptic style and sign choice carry largely independent administrative information, justifying their joint use in the Bayesian chronological model of Appendix B.5 without concern for multicollinearity.

16.4 The Question of Proto-Elamite "State" Formation

One prominent interpretive framework for the proto-Elamite corpus connects it to state formation: the hypothesis that the administrative tablet system served a nascent proto-Elamite state apparatus, analogous to the role of proto-cuneiform in the Uruk IV temple administration. This hypothesis implies, among other things, that the corpus represents a single, centrally maintained administrative technology rather than a polycentric or distributed notation.

© Silvano Salvador

The evidence of the present monograph bears directly on this question, though it does not resolve it.

Note F.2: What the Statistical Evidence Says and Does Not Say About State Formation

Three findings of this monograph are relevant to the state-formation hypothesis. Finding 4.2 (Susa-outward transfer entropy) is consistent with Susa playing a central coordinating role, which would support a hierarchical state interpretation; but transfer entropy alone cannot distinguish centralized political authority from voluntary commercial convention diffusion. Finding 2.1 (PE III sign-repertoire drift) is consistent with the dissolution of a previously centralized administrative network, which has been interpreted by some scholars as evidence for state collapse in PE III; but it is equally consistent with the fragmentation of a non-hierarchical commercial network. Finding 6.4 (rejection of H4, the localized-notation hypothesis, for PE I–II) establishes that the script was *not* purely local during its main phase, which is a necessary but not sufficient condition for state-administered uniformity. In summary, the statistical evidence is compatible with but does not entail the state-formation hypothesis; it is also compatible with the Network Horizon model's polycentric-coordination alternative.

16.5 Material Culture Parallels: Tokens, Bullae, and Numerical Tablets

Before the proto-Elamite script proper, the administrative needs of the emerging Susiana complex were partially served by a system of clay tokens enclosed in sealed clay envelopes (bullae): three-dimensional accounting devices that recorded commodity quantities by the shape and number of enclosed tokens, with the sealed envelope guaranteeing the integrity of the record. The transition from token-in-bulla to two-dimensional clay tablet (with impressed signs representing the formerly three-dimensional tokens) is attested archaeologically at Susa in the late Uruk period, providing a partial material genealogy for the proto-Elamite numerical sub-system. The non-numerical signs, however, have no clear token-system precursor, which is one reason the administrative-semasiography hypothesis (H2) is typologically consistent with their origin: they represent novel inventions needed to specify administrative details (commodity sub-type, institutional source, personnel category) that the token system addressed by physical distinctions among three-dimensional token shapes.

Worked Example F.1: The Bulla-to-Tablet Transition at Susa

Among the 412 stratigraphically-contextualized tablets of the Bayesian chronological model (Proposition 2.1), 37 come from excavation contexts also yielding clay bullae or loose tokens. In 34 of these 37 contexts (91.9%), the bulla/token assemblage and the tablet assemblage are stratigraphically adjacent rather than overlapping, consistent with a transitional period in which both technologies co-existed before the tablet system fully supplanted the bulla system. Only 3 of the 37 contexts show clear stratigraphic intermixture, and all 3 were flagged as disturbance contexts in the original excavation reports (consistent with the PE III fragmentation finding of Chapter 2). This spatial patterning supports the standard interpretation of the bulla-to-tablet transition as a roughly contemporaneous, geographically broad shift rather than a prolonged coexistence, and places the invention of the non-numerical sign sub-system at or very near the beginning of the PE I phase.

© Silvano Salvador

Chapter 17

Comparative Epigraphy: Structural Typology of Early Writing Systems

17.1 A Formal Taxonomy of Early Administrative Notations

The reference corpus $\mathcal{R}$ of Definition 5.1 contains six systems spanning the full range from degenerate tallying to fully phonetic syllabic script. This chapter situates proto-Elamite within a broader comparative epigraphy of early administrative notations, expanding the three-statistic classification of Chapter 5 with additional structural properties and a wider set of comparanda drawn from the global archaeological record.

Definition G.1: Extended Typological Classification

A seven-class typological scheme for early administrative notations is proposed, extending the three-class scheme of Definition 5.2:

- **T1. Pure tallying**: marks denoting only "one unit of something", distinguished at most by the physical position of the mark.
- **T2. Numerical notation**: positional-value systems with an explicit sign inventory for different magnitudes and counting domains.
- **T3. Constrained semasiography**: the H2 system proposed for proto-Elamite; combinatorially structured commodity notation without phonological mediation.
- **T4. Proto-logographic**: systems on the trajectory toward logographic writing, exhibiting partial phonetic complementation; proto-cuneiform in its Uruk III phase.
- **T5. Full logographic**: one sign per morpheme or concept, with systematic phonetic extension; Chinese character writing, early Sumerian cuneiform.
- **T6. Logo-syllabic**: logographic base with syllabic extension; Maya writing, Egyptian hieroglyphics.
- **T7. Syllabic/alphabetic**: predominant or exclusive phonetic encoding; Linear B, the Phoenician abjad, the Greek alphabet.

Proposition G.1: Proto-Elamite Occupies Type T3 or T4 in the Extended Scheme

Mapping the six diagnostic statistics of Note 6.1 onto the seven-class typological scheme of Definition G.1, proto-Elamite's profile ($|\Sigma| = 637$, $H_\infty/H_0 = 0.66$, $\rho_{pos} = 0.38$, $k^* = 1$, cross-entropy distances to $\mathcal{R}_1$–$\mathcal{R}_3$) is most consistent with Type T3 (constrained semasiography) but also compatible with an early transitional position in Type T4 (proto-logographic), since the two types differ primarily in the presence or absence of phonetic complementation, which the corpus-internal statistics of this mono-

graph cannot directly detect. Critically, proto-Elamite's profile is *not* compatible with Types T5–T7 (logographic to alphabetic systems), as these types consistently exhibit $H_\infty/H_0 < 0.50$ and $k^* \geq 2$ in the reference panel, values significantly below and above the observed proto-Elamite figures respectively.

17.2 The Indus Valley Script: A Detailed Structural Comparison

The Indus Valley sign system ($\mathcal{R}_4$, Definition 5.1) was introduced in Chapter 5 and briefly compared to proto-Elamite in Section 8.4. Here the comparison is deepened with a sign-level analysis that parallels the proto-Elamite sign-level analysis of Appendix D.

Proposition G.2: Indus Valley and Proto-Elamite Share a Bipartite Sign-Frequency Structure

Both the Indus Valley corpus and the proto-Elamite non-numerical sub-corpus exhibit a bipartite sign-frequency structure in which a small set of "frame signs" (high-frequency, position-variable, cross-site stable) and a large set of "content signs" (lower-frequency, position-specific, more site-variable) can be distinguished using the (ϕ_s, ν_s) diagnostic pair of Proposition 6.3. For Indus Valley, the analogous "frame sign" set has been independently identified by Parpola (1994) as seal-header signs with likely institutional or ownership functions. The structural parallel between the two corpora — frame signs + content signs, without a fully phonological encoding — is consistent with Type T3 (constrained semasiography) or Type T4 (proto-logographic) for both systems, and with the convergent-development argument of Corollary 8.2.

17.3 Egyptian Administrative Notation and the Hieratic Parallel

A comparandum not included in the main reference corpus $\mathcal{R}$ but relevant to understanding the proto-Elamite context is the early Egyptian administrative notation attested at Abydos (Tomb U-j, approximately contemporaneous with PE I) and in the earliest hieratic administrative papyri (c. 3000–2800 BCE). Egyptian writing is known from external evidence to have been from the beginning at least partially phonetic (royal names recorded using the rebus principle appear in the earliest attested royal serekhs), yet the earliest administrative texts from Abydos show a functional profile — commodity tags with numerical counts and institutional emblems — structurally similar to the proto-Elamite administrative case structure.

Note G.1: The Egyptian Parallel and Its Limits

The Egyptian administrative notation from Abydos Tomb U-j ($n \approx 150$ inscribed objects, substantially smaller than the proto-Elamite corpus) cannot be analyzed using the full six-diagnostic battery of Note 6.1 due to corpus-size limitations. A restricted three-statistic analysis ($|\Sigma|$, H_∞/H_0, ρ_{pos}) yields values consistent with Type T3–T4 in Definition G.1's scheme: $|\Sigma| \approx 150$, $H_\infty/H_0 \approx 0.55$, $\rho_{pos} \approx 0.32$ (all estimated with wide confidence intervals at this corpus size). These values are adjacent to but distinct from the proto-Elamite point estimate, consistent with both systems occupying the T3–T4 boundary region of the extended typological scheme but with Egyptian notation having somewhat more phonetic structure (partly due to the known royal-name rebus usage, which has no confirmed proto-Elamite parallel). The comparison illustrates the

© Silvano Salvador

general tendency of early urban administrative notations to cluster in the T3–T4 region regardless of ultimate trajectory.

17.4 Proto-Sinaitic and the Origins of Alphabetic Writing

The typological scheme of Definition G.1 culminates at Type T7 (syllabic/alphabetic), represented in the main reference corpus by Linear B. The path from early administrative semasiography (T3) to alphabetic writing (T7) traverses several intermediate types and, crucially, required for each transition a specific invention or discovery: the rebus principle (enabling T3→T4), systematic phonetic complementation (enabling T4→T5), and the consonantal abstraction that produced the abjad (enabling T5→T7 for Northwest Semitic). Proto-Elamite, on the present analysis, never made the T3→T4 transition (or made it only partially, if the H1′ hypothesis captures some genuine phonetic structure in the corpus): this is the most fundamental sense in which proto-Elamite is an "undeciphered script" that resists the methods that cracked proto-cuneiform, since those methods depended specifically on the T4 phonetic-complement evidence that is absent from the proto-Elamite record.

Worked Example G.2: Entropy Profiles Across the Typological Trajectory

Computing H_∞/H_0 for representative systems at each type in Definition G.1's scheme (using the same Miller–Madow corrected estimator of Definition 3.1, at matched corpus sizes where possible):

Type	Representative system	H_∞/H_0
T1	Notched tally sticks (Ishango, $n = 300$)	0.91
T2	Capacity system N alone ($\mathcal{N}$)	0.83
T3	Synthetic semasiography ($\mathcal{R}_6$)	0.68
T3–T4	Proto-Elamite ($\mathcal{S}$)	0.66
T4	Proto-cuneiform Uruk IV ($\mathcal{R}_1$ subset)	0.61
T5	Chinese oracle-bone script ($n \approx 10{,}000$)	0.49
T6	Egyptian Middle Kingdom hieratic ($n \approx 15{,}000$)	0.41
T7	Linear B ($\mathcal{R}_3$)	0.27

The monotonic decline of H_∞/H_0 across the typological trajectory is striking, and its interpretation is straightforward: more phonetically-encoded systems exhibit more sequential structure (lower residual entropy conditional on context), because phonological and morphological grammar constrains sign order more tightly than the slot-grammar of an administrative semasiography. Proto-Elamite's value of 0.66 places it squarely in the T3–T4 range, between the synthetic semasiographic control and proto-cuneiform — consistent with the H2-favoring conclusion of Corollary 7.1 and the provisional T3–T4 classification of Proposition G.1.

© Silvano Salvador

Chapter 18

Extended Hypothesis Space: Minority Theories and Their Evaluation

18.1 Beyond the Four Principal Hypotheses

The main hypothesis evaluation of Chapters 6–7 considers four hypotheses (H1–H4) that collectively span the space of interpretations advanced in the scholarly literature on proto-Elamite. This appendix treats four minority hypotheses that have been proposed but that lie outside the four-hypothesis framework, either because they apply to specific subsets of the corpus, because they were advanced in the pre-digital era without quantitative grounding, or because they represent theoretical possibilities not yet seriously entertained in the epigraphic literature but worth evaluating for completeness.

18.1.1 H5: A Ritual or Votive Notation

A small minority of scholars (most prominently Meriggi, 1971–1974) has proposed that some proto-Elamite tablets record votive or ritual transactions rather than purely economic ones. Under this hypothesis, the non-numerical signs in ritual contexts would encode a different type of referential content — deity names, ritual formulae, offering specifications — with potentially different statistical properties from the administrative mainstream of the corpus.

Proposition H.1: H5 Is Not Supported as a Corpus-Wide Account

The corpus-level statistics of Chapters 3–5 show no evidence for a bimodal structure in the conditional entropy or positional regularity profiles that would indicate two structurally distinct sub-populations of tablet types (ritual versus administrative). A two-component Gaussian mixture model fitted to the distribution of per-tablet ρ_{pos} values (using the same EM procedure as the three-phase chronological clustering of Section 2.3) is not preferred by BIC over the one-component model ($\Delta\text{BIC} = +4.1$, favoring the one-component model). This finding does not exclude a small votive subset (it could be too small to produce a detectable second mode), but it precludes H5 as a *global* interpretation of a substantial fraction of the corpus.

18.1.2 H6: A Calendar or Astronomical Notation

A single published proposal (not widely cited in the epigraphic literature) has suggested that proto-Elamite non-numerical signs encode a lunisolar calendar system, with specific sign sequences corresponding to month-names, season-markers, or astronomical events. This hy-

pothesis predicts strong periodicity in sign occurrence that should be detectable as autocorrelation in within-tablet sign sequences.

Proposition H.2: H6 Is Decisively Rejected by the Periodicity Test

Applying a discrete autocorrelation test to within-tablet sign sequences (lag 1 through lag 5, using the Ljung–Box statistic at $\alpha = 0.05$ with Holm–Bonferroni correction), no significant autocorrelation is found beyond lag 1 (the bigram dependency already captured by the transition matrix of Definition 4.1). In particular, no signs exhibit a periodic occurrence pattern at lag 3 (consistent with a three-season cycle), lag 4 (four-season), or lag 12 (consistent with a twelve-month calendar), and the maximum Ljung–Box statistic across all sign-lag combinations ($Q_{\max} = 5.8$, $df = 1$, $p = 0.016$) does not survive correction for the 637 $\times$ 5 = 3,185 simultaneous tests conducted. H6 is accordingly rejected without reservation.

18.1.3 H7: A Merchant-Guild or Kinship Notation

A hypothesis that has occasionally surfaced in archaeologically-oriented discussions of the proto-Elamite corpus is that the non-numerical signs encode information about trade guild membership or kinship group affiliation — that is, that the "commodity" being recorded is not a physical good but a social category. This hypothesis predicts that the non-numerical signs should co-vary with seal impressions (since seals identify individuals or institutions) more strongly than with document types (since social categories should be cross-cutting across grain, livestock, and other accounts).

Proposition H.3: H7 Is Partially Supported but Dominated by H2

The association between glyptic style class and non-numerical sign distribution established in Proposition F.1 (Cramér's $V = 0.14$) is non-trivial but modest, providing limited positive evidence for the guild/kinship hypothesis in the sense that seal identity does correlate with sign choice beyond what phase alone explains. However, the document-type association of Proposition D.2 (Cramér's $V = 0.58$) is substantially stronger, indicating that commodity domain is a far better predictor of sign choice than institutional identity. H7 is therefore not supported as a primary account of the non-numerical sign system but is consistent with the H3 (identifier sub-system) interpretation of Finding 6.3: the 73-sign identifier subset may partially encode institutional or kinship-group affiliations, while the majority of the sign inventory reflects commodity-domain information under H2.

18.1.4 H8: An Emerging Phonetic System at Susa Only

A more theoretically interesting minority hypothesis — not explicitly formulated in the published literature but implied by several comparative-epigraphy arguments (see Corollary 8.1) — is that the Susa corpus specifically, as the primary administrative center, may have begun developing phonetic uses of non-numerical signs (a proto-rebus principle analogous to Uruk III phonetic complementation) even while peripheral-site tablets remained purely semasiographic. Under H8, the Susa corpus would exhibit a higher degree of sequential structure (lower H_∞/H_0, higher k^*) than the Anshan or Sialk sub-corpora, reflecting more extensive phonetic constraint on sign order.

© Silvano Salvador

Proposition H.4: H8 Is Not Supported by Site-Specific Entropy Comparisons

Computing site-specific conditional entropy profiles for Susa ($n = 6{,}400$ tokens), Anshan ($n = 1{,}750$ tokens), and Sialk ($n = 610$ tokens, treated with appropriate caution given the statistical power limitations of Note 10.3), the site-specific entropy plateau ratios are: Susa $H_\infty/H_0 = 0.65$ (95% CI $[0.61, 0.69]$), Anshan $H_\infty/H_0 = 0.67$ ($[0.61, 0.73]$), Sialk $H_\infty/H_0 = 0.70$ ($[0.60, 0.80]$, wide CI reflecting small n). The differences are not statistically significant (pairwise t-tests, all $p > 0.20$), and Susa's entropy plateau is not lower than Anshan's (which would be the H8 prediction). H8 is accordingly not supported at the available corpus size, though the wide Sialk CI precludes a definitive verdict for that peripheral site.

18.2 Why the Minority Hypotheses Matter for Future Research

The evaluation of H5–H8 in this chapter is not merely negative. Even hypotheses that are decisively rejected (H6, astronomical notation) or not supported (H5, H8) provide constraints on future research by narrowing the space of viable interpretations. The partial support for H7 (guild/kinship affiliation in the identifier subset) opens a specific research direction: cross-referencing the 73-sign candidate-identifier set of Finding 6.3 with the independently documented proto-Elamite glyptic corpus (thousands of cylinder-seal impressions in museum collections, many not yet systematically linked to tablet find-contexts) to test whether specific identifier signs co-occur preferentially with specific seal motifs, which would provide the first genuine external validation of a proposed sign function.

© Silvano Salvador

Chapter 19

Thematic Glossary: Key Concepts and Technical Terms

Note I.1: Purpose of This Glossary

This glossary defines, in compact form, the key technical concepts used across the monograph, organized thematically rather than alphabetically. It is intended both as a reference for readers encountering specific terms for the first time and as a conceptual map of the analytical framework connecting the monograph's four analytical axes. Cross-references to the main text are given where a term receives its full technical definition.

19.1 Script and Writing-System Concepts

Definition I.1: Semasiography

A **semasiography** (from Greek *sēmasia*, "meaning", and *graphia*, "writing") is a system of graphic notation that conveys meaning directly, through a mapping from sign to concept or referent, without necessarily mediation through the phonological structure of a spoken language. A semasiographic system may be highly structured and combinatorially rich while remaining non-linguistic in the technical sense. The hypothesis H2 of this monograph proposes that the proto-Elamite non-numerical sign sub-system is a constrained semasiography of this type. The term is used in contrast to **glottography**, which denotes systems that encode spoken language (either syllabically or logographically).

Definition I.2: Sign vs. Grapheme vs. Logograph

In the technical vocabulary of this monograph:

- A **sign** (or sign token) is a single impressed or incised graphic unit on a clay tablet, identified by the OPEP sign-numbering convention (PE-nnn). Signs are the basic units of the corpus.
- A **sign type** is an equivalence class of graphic signs sharing the same conventionally recognized shape, regardless of minor inter-token variation. The 637-sign reduced inventory Σ^*_{PE} is an inventory of sign types.
- A **grapheme** is a sign type in a phonetically-encoding script that corresponds to

a phonological unit (phoneme, syllable, or mora). The term is avoided in this monograph because the existence of proto-Elamite graphemes is precisely what is in dispute.

- A **logograph** is a sign type in a logographic script corresponding to a morpheme or word. The term is used only for the proto-cuneiform and Sumerian cuneiform comparanda, where its applicability is independently established.

Definition I.3: Administrative Case and Sign-Group

Following the corpus formalization of Definition 1.2, an **administrative case** is the basic structural unit of a proto-Elamite tablet: a sequence of sign tokens delimited by incised clay rulings, representing a single transaction entry within the account. A **sign-group** is a sub-unit of a case, typically consisting of a numerical notation followed by one or more non-numerical signs. The case $\supset$ sign-group $\supset$ sign hierarchy is the three-level administrative structure that organizes the corpus and serves as the unit of analysis throughout the sequential-dependency analyses of Chapters 3–4.

Definition I.4: Hapax Legomenon

A **hapax legomenon** (from Greek, "said only once") is a sign type attested exactly once in the corpus. The proto-Elamite non-numerical sub-corpus contains 363 hapax legomena, excluded from the reduced inventory Σ^*_{PE} for frequency-distribution analyses but retained for vocabulary-size calculations. The hapax proportion (363 out of 1,000 attested types = 36.3%) is within the range expected for a corpus of this size and Zipf-law parameters (Heaps' law predicts a hapax proportion of $\approx$ 30–40% at $N = 10{,}000$ tokens with $\alpha_Z \approx 1.3$), and does not indicate any unusual degree of sign-system instability beyond what corpus-size effects alone predict.

19.2 Information-Theoretic Concepts

Definition I.5: Shannon Entropy and Its Interpretations

The **Shannon entropy** $H(X) = -\sum_x p(x) \log_2 p(x)$ of a discrete random variable X with probability distribution p measures the average uncertainty (in bits) about the value of X before it is observed. In the corpus analysis of this monograph, H_0 (zeroth-order entropy) measures the uncertainty about which sign type appears at a randomly-chosen token position, given only the marginal frequency distribution; H_k (conditional entropy) measures the residual uncertainty about the next sign given the previous k signs. The *entropy plateau* phenomenon — that H_k stops decreasing after $k = 1$ for proto-Elamite — is the central diagnostic result of Chapter 3 (Finding 3.3, the restricted-register anomaly).

Definition I.6: Cross-Entropy and Kullback–Leibler Divergence

For two probability distributions P and Q over the same alphabet, **cross-entropy** $H(P, Q) = -\sum_x P(x) \log_2 Q(x)$ measures how many bits are required on average to encode samples from P using a code optimized for Q. Cross-entropy is always $\geq H(P)$ (Shannon entropy), with equality only when $P = Q$. The **Kullback–Leibler**

© Silvano Salvador

divergence $D_{KL}(P\|Q) = H(P,Q) - H(P)$ measures the excess description cost; it is zero only when $P = Q$ and strictly positive otherwise. In Chapter 4, both quantities are used to measure the distributional distance between the proto-Elamite corpus and the proto-cuneiform comparandum (Finding 4.1), with the symmetric sum $D_{KL}(P_{PE}\|P_{PC}) + D_{KL}(P_{PC}\|P_{PE}) = 4.29$ bits quantifying the scale of their distributional difference.

Definition I.7: Transfer Entropy

Transfer entropy from process Y to process X, $T_{Y\to X}$, measures the additional predictive power for X's next state provided by Y's current state, beyond what X's own past provides. Unlike correlation, transfer entropy is directional: $T_{Y\to X} \neq T_{X\to Y}$ in general, making it a tool for detecting asymmetric causal influence. In the context of this monograph (Section 4.4), the directional asymmetry $\Delta T = T_{\text{Susa}\to\text{Anshan}} - T_{\text{Anshan}\to\text{Susa}} = 0.043$ bits ($p = 0.01$) is interpreted as evidence that the Susa scribal tradition influenced the Anshan tradition more than vice versa — a statistical signature of the administrative primacy of Susa in the proto-Elamite network.

Definition I.8: Minimum Description Length

The **Minimum Description Length** (MDL) principle is an approach to model selection based on the intuition that the best model is the one that most compresses the data: a model's quality is measured by the total length of the compressed representation (model parameters + data encoded using the model). In two-part MDL (Appendix B.9), the total description length $L(\text{model}) + L(\text{data}|\text{model})$ is minimized over the hypothesis space. In Chapter 7, the MDL differentials of Finding 7.2 establish that H2 (constrained semasiography) achieves the best compression of the proto-Elamite corpus among the five hypotheses, with the H1 (natural-language) model incurring a catastrophic description-length penalty due to the bigram-sparsity problem described in Proposition 4.2.

Definition I.9: Zipf's Law and the Zipf–Mandelbrot Distribution

Zipf's law states that in a corpus drawn from a natural language, the frequency of a word is approximately inversely proportional to its frequency rank: $f(r) \propto r^{-1}$. The **Zipf–Mandelbrot distribution** generalizes this to $f(r) = C(r+b)^{-\alpha}$, adding a rank offset b and a free exponent α. For natural languages, $\alpha \in [0.9, 1.1]$; for the proto-Elamite non-numerical sub-corpus, the maximum-likelihood estimate is $\hat{\alpha} = 1.31$ (Finding 3.2), significantly above the natural-language range and consistent with the more extreme sign-concentration profile expected under the constrained-semasiography hypothesis H2.

© Silvano Salvador

19.3 Bayesian and Statistical Concepts

Definition I.10: Bayes Factor

A **Bayes factor** $K_{ij} = \hat{\mathcal{L}}(H_i)/\hat{\mathcal{L}}(H_j)$ is the ratio of marginal likelihoods of two hypotheses H_i and H_j, quantifying the evidence that the data provide in favor of H_i over H_j independently of prior probabilities. On the Jeffreys/Kass–Raftery scale: $K_{ij} > 3.2$ is "substantial" evidence, $K_{ij} > 10$ is "strong", $K_{ij} > 100$ is "decisive". In Chapter 7, $K_{H2,H1} = 24.1$ is rated "strong to decisive" evidence for H2 over unrestricted H1, while $K_{H2,H1'} = 5.26$ is "strong" evidence for H2 over the restricted-register natural-language variant.

Definition I.11: Jensen–Shannon Divergence

The **Jensen–Shannon divergence** $\mathrm{JSD}(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)$, where $M = (P+Q)/2$ is the mixture distribution, is a symmetric, bounded ($0 \leq \mathrm{JSD} \leq 1$ bit) measure of distributional distance between two probability distributions. Unlike D_{KL}, it is defined even when one distribution assigns zero probability to points in the other's support. In Chapter 2 (Finding 2.1), JSD divergences between PE I, PE II, and PE III unigram distributions are used to measure phase-to-phase sign-frequency drift; the three-fold increase in JSD from PE I→PE II (0.031 bits) to PE II→PE III (0.094 bits) is the primary quantitative evidence for PE III fragmentation.

Definition I.12: Bootstrap Confidence Interval

A **bootstrap confidence interval** is a confidence interval computed by resampling the observed data (with replacement) a large number of times ($B = 10{,}000$ in this monograph), computing the statistic of interest on each resample, and using the resulting empirical distribution of the statistic to set the confidence bounds. The bias-corrected and accelerated (BCa) bootstrap (Appendix B.2) corrects for both the sampling bias and the skewness of the bootstrap distribution, providing more accurate coverage than the simpler percentile bootstrap for skewed statistics such as the Zipf exponent and the entropy plateau ratio.

19.4 Archaeological and Epigraphy Concepts

Definition I.13: Stratigraphy and Stratigraphic Context

Stratigraphy refers to the study of the sequence and vertical arrangement of cultural layers (strata) in an archaeological site, which provides relative chronological ordering of finds. A **stratigraphic context** specifies which excavated layer or locus a tablet was found in. In the proto-Elamite case, the French Délégation Archéologique en Iran excavations at Susa recorded stratigraphic contexts for ≈ 412 tablets (Susa levels XIV–XVII), providing the primary relative dating evidence for the Bayesian chronological model of Chapter 2.

© Silvano Salvador

Definition I.14: Glyptic and Cylinder Seal

Glyptic refers to the art of carving seal impressions, and by extension to the corpus of seal impressions themselves. A **cylinder seal** is a small cylindrical stone or other material carved with a decorative or administrative design that, when rolled across wet clay, leaves a continuous impressed image. Cylinder seals served as personal or institutional identifiers in proto-Elamite administration and are found impressed on both clay tablets and clay sealings. The stylistic evolution of cylinder-seal designs across the late Uruk to Jemdet Nasr period provides the chronological seriation used in the glyptic likelihood function of Definition 2.2.

Definition I.15: Bulla and the Envelope System

A **bulla** (plural: bullae) is a hollow clay envelope used in pre-tablet accounting systems: clay tokens representing commodity units were enclosed in the bulla and the envelope was sealed to guarantee the integrity of the enclosed tokens. The outer surface of the bulla was sometimes marked with impressions of the enclosed tokens to allow the quantity to be read without breaking the seal. The proto-Elamite tablet system replaced the bulla system during the PE I phase, as documented in Worked Example F.1 and the associated archaeological literature.

Definition I.16: Cuneiform and Cuneus

Cuneiform (from Latin *cuneus*, "wedge") refers to the family of ancient Near Eastern writing systems that used a reed stylus pressed into clay to produce characteristic wedge-shaped impressions. Proto-Elamite is technically a cuneiform-adjacent script in that it uses a similar reed-stylus-on-clay medium, but its sign forms are not the triangular wedge combinations characteristic of fully-developed cuneiform writing; instead, proto-Elamite signs use both impressed and incised strokes, a mixed technique that distinguishes the medium from both proto-cuneiform and later cuneiform scripts. The numerical notations of both proto-cuneiform and proto-Elamite are, however, impressed with circular or semi-circular stylus tips, producing the shared "numerical notation" appearance noted in Section 8.1.

19.5 Extended Sign Catalogue: Frequency and Distributional Data

Note I.2: Scope of the Extended Sign Catalogue

The following tables provide frequency and distributional data for representative signs across the four frequency classes of Definition D.1. Full per-sign data for all 637 types in Σ^*_{PE} are available in the project's computational data release (see Appendix B.10 for the computational environment); the tables here provide a stratified sample of 80 signs (20 per class) intended to illustrate the within-class distributional patterns discussed in Appendix D.

© Silvano Salvador

Table 19.1: Class I Signs: Top 20 by Frequency Rank

Sign ID	OPEP Ref.	Tokens	$\hat{p}$ (%)	ϕ_s	JSD (SUS/ANS)
PE-001	M001	487	4.80	0.94	0.03
PE-002	M002	315	3.11	0.32	0.04
PE-003	M003	274	2.70	0.41	0.05
PE-004	M004	248	2.45	0.28	0.03
PE-005	M005	221	2.18	0.35	0.04
PE-006	M006	203	2.00	0.27	0.04
PE-007	M007	188	1.85	0.31	0.05
PE-008	M008	176	1.74	0.38	0.05
PE-009	M009	161	1.59	0.26	0.04
PE-010	M010	148	1.46	0.33	0.04
PE-011	M011	139	1.37	0.29	0.05
PE-012	M012	131	1.29	0.30	0.05
PE-013	M013	124	1.22	0.27	0.04
PE-014	M014	117	1.15	0.31	0.05
PE-015	M015	112	1.10	0.34	0.06
PE-016	M016	106	1.05	0.28	0.05
PE-017	M017	101	1.00	0.32	0.06
PE-018	M018	97	0.96	0.35	0.05
PE-019	M019	93	0.92	0.30	0.06
PE-020	M020	89	0.88	0.33	0.07
Class I total		3,531	34.8%	0.33	0.047

Note: ϕ_s = positional fixity index; JSD (SUS/ANS) = Jensen–Shannon divergence between Susa and Anshan site-specific frequency distributions. All Class I signs have JSD < 0.10 bits, confirming their cross-site stability (Proposition D.1).

Table 19.2: Class IV Signs: A Stratified Sample of 20 Rare Signs

Sign ID	OPEP Ref.	Tokens	$\hat{p}$ (%)	ϕ_s	Sites attested
PE-401	M401	5	0.049	0.60	SUS only
PE-412	M412	4	0.039	0.75	SUS, ANS
PE-425	M425	4	0.039	0.50	ANS only
PE-438	M438	3	0.030	0.67	SUS only
PE-444	M444	3	0.030	0.67	SUS only
PE-451	M451	3	0.030	1.00	SIA only
PE-463	M463	3	0.030	0.67	SUS, YAH
PE-477	M477	2	0.020	0.50	ANS only
PE-489	M489	2	0.020	1.00	SUS only
PE-501	M501	2	0.020	0.50	SUS only
PE-514	M514	2	0.020	0.50	SUS, ANS
PE-522	M522	2	0.020	1.00	YAH only
PE-536	M536	2	0.020	0.50	SUS only
PE-548	M548	2	0.020	1.00	ANS only

© Silvano Salvador

Table 19.2: Class IV Signs: A Stratified Sample of 20 Rare Signs

Sign ID	OPEP Ref.	Tokens	$\hat{p}$ (%)	ϕ_s	Sites attested
PE-561	M561	2	0.020	0.50	SUS only
PE-574	M574	2	0.020	0.50	SUS, SIA
PE-585	M585	2	0.020	0.50	SUS only
PE-599	M599	2	0.020	1.00	SUS only
PE-611	M611	2	0.020	0.50	ANS only
PE-624	M624	2	0.020	0.50	SUS only
Class IV sample total		51	0.50%	0.65	mostly 1 site

Note: The high proportion of single-site attestation in Class IV (14 of 20 sampled signs = 70%, compared to 0% for Class I) directly illustrates Proposition D.3's site-specificity finding. The elevated mean $\phi_s = 0.65$ for rare signs reflects the fact that a sign attested only 2–5 times will have most of those occurrences at a single case-position by chance.

19.6 The Slot Grammar in Detail: A Formal Specification

Definition I.17: The Proto-Elamite Slot Grammar $\mathcal{G}_{PE}$

Based on the distributional analyses of Chapters 4–5 and Appendix D, the constrained-semasiography hypothesis H2 implies the existence of a slot grammar $\mathcal{G}_{PE}$ governing case-internal sign order. A minimal formal specification consistent with the observed positional regularity ($\rho_{pos} = 0.38$), effective model-order ceiling ($k^* = 1$), and community structure (Proposition D.4) has the following structure:

1. **Slot 1 (Domain classifier):** Mandatory, case-initial; draws from Community 1 or 2 of G_{PE}'s Louvain communities; predominantly Class I signs.
2. **Slot 2 (Commodity sub-type):** Mandatory, medial; draws from the community corresponding to the Domain classifier in Slot 1; Class I–III signs.
3. **Slot 3 (Quantity qualifier):** Optional, medial; draws from a small vocabulary of quantity-modifier signs; Class II signs.
4. **Slot 4 (Institution/personnel identifier):** Optional, case-final; draws from the 73-sign candidate-identifier subset (Finding 6.3); predominantly Class II–III signs.
5. **Slot 5 (Domain re-classifier or case-close marker):** Optional, final; predominantly PE-001 (the high-frequency case-framing sign of Case Study 1).

Under this specification, a minimal case has 2 sign tokens (Slots 1+2) and a maximal case has 5 sign tokens (all five slots), consistent with the observed case-length distribution (mean 4.3 tokens, standard deviation 1.7).

Proposition I.1: The Slot Grammar Accounts for 78% of Observed Sign Combinations

Applying the slot-grammar specification of Definition I.17 to the full non-numerical sub-corpus, 78% of administrative cases are consistent with the five-slot template (i.e., every sign token in the case can be assigned to a slot with the correct community membership and positional constraints). The remaining 22% of cases include at least

© Silvano Salvador

one sign token that either fills an unspecified position, combines community memberships in a way not predicted by the template, or uses a sign type not unambiguously assignable to any slot by frequency class alone. This 78% coverage rate is consistent with the H2 hypothesis's prediction that the slot grammar is "constrained but not rigid": administrative practice in a real-world setting exhibits enough flexibility to accommodate exceptional commodity combinations, institutional contexts, and scribal conventions not fully captured by the minimal five-slot model.

© Silvano Salvador

Chapter 20

A Historiography of Proto-Elamite Scholarship

20.1 Early Discovery and the Susa Excavations, 1897–1905

The Proto-Elamite corpus entered scholarly awareness through Jacques de Morgan's excavations at Susa beginning in 1897, conducted under the auspices of the French Délégation en Perse. De Morgan's teams, working with the recording standards of their era, recovered the bulk of what is today catalogued as the "Acropole" tablet deposit without the stratigraphic precision that later excavators would demand. This circumstance has a direct and persistent consequence for the quantitative analysis pursued throughout this monograph: a substantial fraction of the corpus — on the order of forty percent of Susa tablets by current estimate — carries find-context information too coarse to support fine-grained phase assignment, and the three-phase chronology of Chapter 1 must therefore be calibrated against a minority of well-documented finds and extrapolated, with appropriate uncertainty, to the remainder.

Historiographical Caution

Any entropy or frequency statistic computed "by phase" inherits the chronological uncertainty of its constituent tablets. Where this monograph reports phase-specific quantities, the reader should understand these as estimates conditioned on the subset of tablets with adequate stratigraphic documentation, not as population parameters for the corpus as a whole.

The initial publication of Proto-Elamite material, by Vincent Scheil in the *Mémoires de la Délégation en Perse* (1900–1905), presented hand-copies of selected tablets without a systematic sign list. Scheil's copies, while historically valuable, are known to contain transcription inconsistencies that modern collation against photographs has only partially resolved. This is a second source of irreducible noise in any large-scale computational study: a small fraction of signs in the published record reflect copyist interpretation rather than directly observed surface forms.

20.2 The Sign-List Era: Scheil, Friberg, and the Search for a Standard Inventory

The absence of a stable, universally adopted sign list constrained Proto-Elamite studies for most of the twentieth century. Successive scholars produced overlapping but non-identical sign inventories, with sign counts ranging from under three hundred to over fifteen hundred

depending on whether allographic variants, damaged or ambiguous forms, and numerical notation were counted separately or merged. The sign list employed in this monograph follows the consolidated inventory developed at the Cuneiform Digital Library Initiative (CDLI) and subsequently refined by the DEPCAS (Digital Exploration of Proto-Elamite and Comparable Administrative Scripts) project, comprising approximately 1,260 distinct non-numerical signs after allograph merging.

Table 20.1: Major Proto-Elamite sign-list editions and their reported sign counts

Edition	Approx. sign count	Status
Scheil (1905)	∼200	Partial, no systematic numbering
Friberg (1978–79)	∼420	Focused on numerical notation
Damerow & Englund (1989)	∼1,000	First near-comprehensive list
Dahl (2005, unpub.)	∼1,050	Refinement of allograph mergers
CDLI/DEPCAS (current)	∼1,260	Basis for this monograph

Each revision of the sign list changes, sometimes substantially, the entropy and Zipfian statistics reported by earlier studies, since merging or splitting allographs directly alters the effective vocabulary size V entering the entropy estimators of Chapter 2. The robustness analysis of Chapter 21 quantifies how sensitive the headline entropy estimates of this monograph are to plausible alternative sign-list choices.

20.3 The Numerical Breakthrough: Friberg and Proto-Elamite Arithmetic

The single most secure decipherment achievement in Proto-Elamite studies belongs not to the main sign system but to its numerical notation. Jöran Friberg's work in the late 1970s, building on earlier observations by Scheil, established that Proto-Elamite numerical signs encode at least two, and probably more, distinct metrological systems: a sexagesimal-adjacent system for discrete counting and one or more bespoke systems for grain and area measurement, with internal base structures of 5, 6, and 10 appearing in different combinations across contexts. This achievement, decades ahead of any comparable progress on the non-numerical sign system, supplies one of the few points of genuinely secure semantic ground-truth available to the present study, and is discussed in its quantitative implications in Chapter 22.

20.4 The Stalled Decades: 1980–2000

Following Friberg's numerical breakthrough, the 1980s and 1990s saw comparatively little progress on the non-numerical sign system. Several factors converged: the limited size of the published corpus relative to the labor required for systematic sign-by-sign analysis; the absence of a bilingual or otherwise externally anchored text of any kind; and a broader scholarly pessimism, articulated by several contributors to the 1996 *World's Writing Systems* volume, regarding the prospects for further advance without new textual discoveries. Englund's 1996 survey chapter, frequently cited in this monograph, represents the most influential statement of this position and remains a touchstone for the constrained-semasiography hypothesis (H2) evaluated in Chapter 5.

© Silvano Salvador

20.5 The Digital Turn: CDLI and the 2000s Revival

The establishment of the Cuneiform Digital Library Initiative in 2000, and its subsequent extension to cover non-cuneiform early writing systems including Proto-Elamite, transformed the empirical basis for the field by making systematically transcribed, photographically documented tablet records available in machine-readable form. This digitization effort is the direct precondition for every quantitative analysis in the present monograph: without a corpus of comparable scale and consistency to the roughly 1,600 tablets analyzed here, none of the entropy, n-gram, or Bayesian model-comparison results of Chapters 2 through 6 could be computed with any confidence.

Methodological Continuity

The quantitative program of this monograph builds directly on, rather than supersedes, three strands of prior work: Friberg's numerical decipherment (treated as ground truth where relevant), Damerow and Englund's sign-list and slot-structure observations (formalized in the slot grammar of Section 14.6), and Dahl's documentation of scribal-hand variation (extended quantitatively in Section 2.5). The contribution of this monograph is not a new data source but a new analytical lens applied to an existing, carefully curated evidentiary base.

20.6 Recent Developments and Unresolved Controversies

The past fifteen years have seen a marked increase in computational approaches to Proto-Elamite, mirroring a broader trend across undeciphered-script studies. Three areas of active disagreement persist and bear directly on the hypothesis-evaluation program of this monograph:

1. **The status of the restricted-register anomaly** (Section 2.3): whether the markedly low positional entropy in certain tablet zones reflects a genuine structural property of the underlying notation or an artifact of corpus composition skewed toward a narrow range of administrative transaction types.

2. **The linguistic affiliation question**: whether, if Proto-Elamite does encode a natural language, that language bears any demonstrable relationship to Elamite as attested in the much later Linear Elamite and Old Elamite cuneiform records (Chapter 7), a question on which specialists remain divided and on which this monograph deliberately declines to take a strong position beyond the Bayesian posterior reported in Chapter 6.

3. **The comparability of Proto-Elamite to Indus Valley notation** (Section 7.4): a debate inherited largely from typological arguments rather than direct contact evidence, given the geographic distance between the two systems, but kept alive by structural similarities in sign repetition and lack of narrative syntax.

This monograph does not resolve these controversies. Its contribution is to supply a transparent, reproducible quantitative framework within which competing positions on each question can be evaluated against a common evidentiary standard, and to report, where the evidence permits, a calibrated degree of support for each position rather than a categorical verdict.

© Silvano Salvador

Chapter 21

Robustness, Simulation, and Sensitivity Analysis

21.1 Motivation: Why Robustness Checks Are Indispensable for a Small, Heterogeneous Corpus

Every quantitative result reported in Chapters 2 through 6 of this monograph rests on a corpus that is small by the standards of computational linguistics (on the order of 1.7×10^4 sign tokens), heterogeneous in documentation quality across sites and phases, and subject to several non-trivial preprocessing choices — sign-list edition, allograph merging policy, treatment of damaged signs, and phase assignment for ambiguously dated tablets. A central methodological commitment of this monograph is that no headline statistic should be reported without an accompanying assessment of its sensitivity to these choices. This chapter consolidates the robustness program, referenced piecemeal in earlier chapters, into a single systematic treatment.

21.2 Perturbation Design: A Taxonomy of Analytic Choices

Definition: Perturbation Axis

A *perturbation axis* is a discrete or continuous analytic choice — such as which sign-list edition to adopt, or which damaged signs to exclude — whose variation defines a family of alternative analyses against which the robustness of a headline result is assessed.

Five perturbation axes are considered systematically in this chapter.

21.3 Bootstrap and Jackknife Estimates of Headline Statistics

For each headline statistic $\hat{\theta}$ reported in Chapters 2 through 6, a non-parametric bootstrap distribution is constructed by resampling tablets with replacement, $B = 2{,}000$ times, recomputing $\hat{\theta}$ on each resample, and reporting the resulting empirical 95% interval. This procedure respects the tablet, rather than the sign token, as the unit of resampling, since signs within a tablet are not independent draws and treating them as such would understate the true sampling variability.

The unigram entropy estimate is comparatively tightly bounded, reflecting its dependence on the full corpus; positional entropy at individual slots, computed on a much smaller

Table 21.1: Perturbation axes examined in the robustness program

Axis	Range examined	Principally affected statistics
Sign-list edition	Damerow–Englund through current CDLI/DEPCAS	Vocabulary size, entropy, Zipf exponent
Damaged-sign treatment	Exclude / impute by context / impute by frequency	Entropy, n-gram models
Phase assignment	Strict (documented only) vs. extrapolated	Diachronic trend estimates
Corpus subsampling	Bootstrap, jackknife by site	All headline statistics
Tokenization of compound signs	Split vs. merge	Vocabulary size, slot-grammar coverage

Table 21.2: Bootstrap 95% intervals for selected headline statistics

Statistic	Point estimate	Bootstrap 95% CI	Relative width
H_1 (unigram entropy, bits)	7.92	[7.71, 8.10]	4.9%
Zipf exponent α	1.07	[0.98, 1.17]	17.8%
Positional entropy, Slot 3	3.14	[2.81, 3.49]	21.7%
Cross-entropy vs. proto-cuneiform	9.85	[9.41, 10.28]	8.8%

effective sample, is correspondingly less precise, and the discussion of the restricted-register anomaly in Section 2.3 should be read with this wider uncertainty band in mind.

21.4 Sign-List Sensitivity: Re-Running the Core Analyses Under Alternative Inventories

Re-running the unigram and bigram entropy computations under each of the five sign-list editions in Table 16.1 produces estimates that vary by no more than 0.4 bits for unigram entropy and no more than 0.6 bits for conditional bigram entropy, a range small enough that no qualitative conclusion of Chapters 2–3 changes under any historically attested sign-list choice. The Zipf exponent is more sensitive, varying between 0.91 and 1.14 depending on whether rare allographic variants are merged aggressively or conservatively, since merging disproportionately inflates the frequency of already-common signs.

Interpretive Note

The sign-list sensitivity result supports treating the qualitative finding of "markedly sub-Zipfian or near-Zipfian behavior with a flattened low-frequency tail" (Section 2.2) as robust, while treating any specific point estimate of α as provisional pending further consolidation of the sign list.

© Silvano Salvador

21.5 Simulated Null Corpora: Calibrating Statistical Significance

To assess whether observed deviations from a null model (Section 2.2's discussion of Zipfian fit; Section 4.5's synthetic semasiography control) could plausibly arise from sampling noise alone given the corpus's modest size, this chapter constructs $1{,}000$ simulated corpora of matched size under three null generative processes: (i) an i.i.d. multinomial draw from the empirical unigram frequency distribution, (ii) a first-order Markov process fitted to the observed bigram transition matrix, and (iii) a strict Zipfian generator with exponent fit by maximum likelihood to the observed frequency-rank data. For each simulated corpus, the same battery of statistics — entropy, positional entropy by slot, and the Bayes factor computation of Chapter 6 — is recomputed, yielding an empirical null distribution against which the observed corpus statistic is compared.

The observed positional entropy at Slot 3 falls below the 1st percentile of the null distribution generated under model (ii), supporting the claim, advanced in Section 2.3, that the restricted-register anomaly is unlikely to be a statistical artifact of corpus size given even a first-order Markov null that already accounts for local sequential dependence.

21.6 Subsampling by Site: Is Susa Driving the Result?

Because Susa supplies approximately 78% of the corpus by tablet count, a natural concern is that any corpus-wide statistic may simply reflect Susa administrative practice and not generalize to the smaller Anshan, Sialk, and peripheral-site assemblages. Recomputing the unigram entropy, Zipf exponent, and restricted-register positional-entropy profile on the Susa-excluded subcorpus (the remaining 22% of tablets) yields entropy and Zipf estimates within the bootstrap confidence intervals of Table 16.2, and a restricted-register effect of the same sign and comparable, though noisier, magnitude. This result supports treating the restricted-register anomaly as a property of the Proto-Elamite notational system generally rather than of Susa scribal practice specifically, though the reduced sample size of the non-Susa subcorpus widens all confidence intervals substantially and this conclusion should be regarded as provisional.

21.7 Summary: A Robustness Scorecard

Table 21.3: Robustness scorecard for headline findings of this monograph

Finding	Robustness	Principal residual uncertainty
Sub-/near-Zipfian frequency distribution	High	Allograph merging policy
Restricted-register positional anomaly	High	Sample size at individual slots
H2/H4 posterior dominance (Chapter 6)	Moderate	Prior elicitation sensitivity (Section 6.5)
Diachronic entropy stability across phases	Moderate	Phase-assignment uncertainty
Cross-entropy gap vs. proto-cuneiform	High	Reference-corpus comparability

© Silvano Salvador

The overall picture that emerges from this robustness program is one of qualitative stability under reasonable perturbation of analytic choices, combined with appropriately wide uncertainty on quantitative point estimates, particularly those computed on small sub-populations of the corpus such as individual slots, phases, or peripheral sites.

© Silvano Salvador

Chapter 22

Metrological Systems: A Formal Reconstruction of Proto-Elamite Numeracy

22.1 The Numerical Sign System as Ground Truth

Among the many open questions surrounding Proto-Elamite, the structure of its numerical notation is exceptional in having been substantially resolved. Friberg's reconstruction, refined by subsequent work, identifies a family of at least four distinct metrological systems operating in parallel within the corpus, distinguished by the base structure of their numeral signs and by the commodity classes with which each system co-occurs.

Definition: Metrological System

A *metrological system* is a closed set of numeral signs $\{n_1, \ldots, n_k\}$ together with a fixed sequence of conversion ratios $r_i = n_{i+1}/n_i$ relating each sign to its successor, used consistently to quantify a particular class of commodity or unit of account.

Table 22.1: Principal Proto-Elamite metrological systems (after Friberg)

System	Base structure	Primary domain
Sexagesimal (S)	1–10–60–600–3600	Discrete counting (animals, people)
Bisexagesimal (B)	1–6–60–...	Grain capacity, alternate accounting
System C	1–10–... irregular	Specific commodity classes
Area system	Distinct unit progression	Land measurement

22.2 Why the Numerical System Decipherment Constrains, but Does Not Resolve, the Main Script Question

The security of the numerical decipherment provides an important methodological anchor: it demonstrates that Proto-Elamite scribes operated a genuinely systematic, internally consistent notational apparatus capable of representing multi-system quantitative information with precision, which weighs against any account of the main sign system as wholly unsystematic ad hoc notation. At the same time, the numerical system's interpretability rests heavily on

its structural regularity and on cross-tablet arithmetic consistency — a tablet's stated total must equal the sum of its line items, a constraint Friberg exploited directly. The main non-numerical sign system offers no comparable internal arithmetic check, which is precisely why its decipherment has proven so much more resistant.

22.3 Quantitative Analysis of Numeral Co-occurrence with Commodity Signs

This section treats the numerical system not as solved background but as a further data source: examining which commodity signs co-occur with which metrological system yields a partition of the commodity vocabulary that is independently informative about administrative semantics, since commodities sharing a metrological system are likely to share a unit of account and plausibly a broader administrative category.

Worked Example: System Partition of Commodity Signs

Restricting to commodity signs (Class II of Section 12.3) that co-occur with a metrological system in at least 20 tablets, a clear bipartition emerges: a "grain cluster" of 14 signs co-occurring overwhelmingly (>85% of instances) with System B, and a "discrete-goods cluster" of 23 signs co-occurring overwhelmingly with System S. A residual group of 6 signs shows mixed co-occurrence, plausibly reflecting commodities measured differently across administrative contexts or sites.

22.4 Metrological Evidence Bearing on the Slot Grammar

The five-slot administrative grammar formalized in Section 14.6 receives additional support from the metrological analysis: the slot identified there as the "quantity slot" is, without exception in the sampled tablets, filled by a sign drawn from one of the four metrological systems catalogued in Table 17.1, and never by a sign from the general (non-numerical) inventory. This provides an independent, structurally motivated criterion for slot identification that does not rely on positional-frequency statistics alone.

22.5 Cross-System Comparison: Proto-Elamite Numeracy Against Proto-Cuneiform

Proto-cuneiform numerical notation, by contrast, is known to have employed an even larger number of context-dependent metrological systems — on the order of a dozen, by Nissen, Damerow, and Englund's count — a complexity that Friberg and others have read as evidence of an administrative environment managing a wider range of distinct commodity classes at Uruk than at contemporary Susa. Whether this reflects genuine economic complexity or simply differential preservation and publication remains open, but the comparison is a useful caution against over-interpreting Proto-Elamite's somewhat simpler numeral apparatus as evidence of a correspondingly simpler administrative system overall.

© Silvano Salvador

Chapter 23

Missing-Data Imputation and Sign-Frequency Forecasting

23.1 The Missing-Data Problem in a Damaged Corpus

A substantial fraction of Proto-Elamite tablets are broken, abraded, or otherwise incompletely legible, and the resulting missing or ambiguous sign readings are not missing at random: damage correlates with tablet size, findspot conditions, and, to some degree, with administrative document category, since larger, more heavily handled administrative summaries appear disproportionately represented among well-preserved finds. Any statistic computed only on fully legible signs risks a systematic bias whose direction is not obvious *a priori*.

23.2 A Multiple-Imputation Framework

Definition: Multiple Imputation for Sign Sequences

Given a damaged sign position with k candidate readings consistent with the surviving trace evidence, multiple imputation draws M complete pseudo-corpora by sampling a candidate reading at each damaged position according to a proposal distribution informed by local context (the n-gram models of Chapter 3) and by the global sign-frequency distribution, computes the target statistic on each pseudo-corpus, and reports the statistic's mean and between-imputation variance across the M draws as an estimate of the additional uncertainty contributed by the missing-data problem.

With $M = 500$ imputations, the unigram entropy estimate increases by 0.06 bits on average relative to the legible-signs-only estimate, with an additional variance component contributing roughly 15% to the total reported uncertainty in Table 16.2. This indicates that missing data is a real but secondary contributor to overall estimation uncertainty, smaller than corpus-size sampling variability but not negligible.

23.3 Forecasting the Undiscovered Tail: A Capture–Recapture Approach

A separate question concerns how much of the true Proto-Elamite sign inventory remains undocumented because it occurs too rarely to have appeared in the excavated and published corpus to date. Treating each tablet as a "capture occasion" and applying a Chao-type nonparametric richness estimator to the sign-frequency table yields an estimated total inventory

size of $\hat{V}_{\text{total}} \approx 1{,}340$, compared to the $1{,}260$ signs actually attested, suggesting that on the order of 80 additional, very low-frequency signs likely exist in the underlying (excavated but unpublished, or unexcavated) population and have simply not yet been observed.

Implication for Decipherment Prospects

An estimated undiscovered tail of roughly 6% of the total inventory is modest and does not, by itself, suggest that the corpus is too small to support meaningful structural analysis of the attested system. It does suggest caution in treating the lowest-frequency attested signs as a closed set: several may yet be merged with currently distinct catalog entries as further tablets are published.

© Silvano Salvador

Chapter 24

A Decision-Theoretic Framework for Evaluating Future Decipherment Claims

24.1 The Recurring Problem of Premature Decipherment Announcements

The history of undeciphered-script scholarship, Proto-Elamite included, is punctuated by claimed decipherments that failed to achieve consensus acceptance, typically because the proposed reading could not be shown to generalize beyond the small set of signs or tablets motivating it. This chapter proposes an explicit decision-theoretic protocol, building on the Bayesian and MDL machinery of Chapter 6, for evaluating the evidentiary weight that a future decipherment claim ought to be assigned before being incorporated into the scholarly consensus.

24.2 A Formal Criterion: Held-Out Predictive Performance

Definition: Held-Out Decipherment Test

Given a proposed decipherment D assigning semantic or phonetic values to a subset of signs, partition the corpus into a training set $\mathcal{T}$ used to calibrate any free parameters of D, and a held-out set $\mathcal{H}$ disjoint from $\mathcal{T}$. The held-out predictive log-likelihood $\ell(D) = \sum_{t \in \mathcal{H}} \log P(t \mid D)$ is computed under the decipherment's implied generative model, and compared against the same quantity under the best-performing null model of Chapter 4 (the Kneser–Ney n-gram baseline). A decipherment claim earns provisional credibility only if $\ell(D)$ exceeds the null baseline by a margin large enough to survive the permutation-test correction of Section 15.3.

This criterion deliberately sets a high bar: a decipherment that merely reproduces, sign by sign, patterns already captured by an unsupervised n-gram model contributes no information beyond what is already known, and should not be credited as a structural advance regardless of its intuitive plausibility or its proponents' philological credentials.

24.3 Decision Costs: False Acceptance Versus False Rejection

Worked Example: An Asymmetric Loss Function

Let L_{FA} denote the cost of provisionally accepting a decipherment claim that later proves spurious (wasted research effort building on a false foundation, reputational cost to the field), and L_{FR} the cost of rejecting a claim that later proves correct (delayed progress). Historical experience across multiple undeciphered scripts suggests $L_{\mathrm{FA}} \gg L_{\mathrm{FR}}$ for scripts with a small, finite tablet corpus, since false acceptance can misdirect an entire subsequent literature in a field too small to easily self-correct through replication. This asymmetry motivates the conservative threshold structure of the held-out test above.

24.4 Applying the Framework Retrospectively

Applying the held-out predictive criterion retrospectively to several historical Proto-Elamite sign-reading proposals (not separately decomposed here to avoid relitigating specific scholars' work) shows that the numerical-system decipherment of Friberg comfortably clears the bar — its predictive advantage over a null model is overwhelming and holds across every held-out partition tested — while no non-numerical sign-reading proposal surveyed achieves a statistically robust advantage. This is consistent with, and provides an independent quantitative gloss on, the qualitative scholarly consensus that the non-numerical system remains undeciphered.

© Silvano Salvador

Chapter 25

Spatial Statistics of Tablet Find-Contexts Within Sites

25.1 Intra-Site Spatial Clustering of Administrative Categories

Beyond the inter-site network analysis of Section 1.10, this chapter examines spatial clustering of tablet content *within* the best-documented find-context, the Susa Acropole sounding, where excavation records permit approximate locus-level assignment for a subset of tablets.

Table 25.1: Spatial association statistics, Susa Acropole sounding (n = 412 tablets with locus data)

Administrative category	Moran's I	p-value (permutation)
Grain accounts	0.31	<0.01
Personnel/labor accounts	0.18	0.04
Animal accounts	0.09	0.21
Mixed/summary accounts	0.22	0.02

The significant positive spatial autocorrelation for grain accounts is consistent with a model in which administrative archives were maintained by locus-specific functional units (e.g., a granary-adjacent record-keeping point) rather than centrally pooled and only subsequently dispersed by post-depositional processes, though the coarse spatial resolution of the available locus data limits the strength of this inference considerably.

25.2 Depositional Process and the Risk of Spurious Spatial Pattern

A caution applies throughout: the Acropole sounding represents a secondary depositional context (most plausibly a discard or archive-clearance deposit rather than a primary administrative findspot), and spatial association in such contexts may reflect deposition order and excavation-trench geometry as much as original administrative organization. The Moran's I statistics reported above should accordingly be read as suggestive of, rather than conclusive evidence for, functionally organized record-keeping loci.

Chapter 26

Formal Models of Scribal Error and Orthographic Variation

26.1 Distinguishing Error from Variation

A persistent interpretive difficulty in Proto-Elamite studies is distinguishing genuine scribal error — a sign written in a form or position inconsistent with the scribe's own evident intention, as judged by internal arithmetic inconsistency or anomalous sign sequences — from legitimate orthographic variation, in which multiple surface forms are equally valid encodings of the same underlying notational unit. The sign-list construction discussed in Section 16.2 must resolve this distinction at the level of individual sign tokens, and the choice made there propagates through every subsequent analysis.

Definition: Arithmetic Error Signal

A tablet exhibits an *arithmetic error signal* when the sum of its itemized numerical entries, computed under the relevant metrological system (Chapter 22), fails to equal its stated total by an amount not attributable to a plausible reading ambiguity in either the items or the total.

26.2 A Catalogue of Detected Arithmetic Discrepancies

Applying this criterion across the subset of the corpus with both itemized entries and an explicit total (approximately 340 tablets), 27 tablets (7.9%) exhibit an arithmetic error signal exceeding the threshold for plausible reading ambiguity. This rate is broadly comparable to error rates documented in contemporary proto-cuneiform administrative accounts, where Englund reports comparable discrepancy frequencies, suggesting that Proto-Elamite scribal practice was not unusually error-prone relative to its closest comparandum.

Table 26.1: Distribution of detected arithmetic discrepancies by site

Site	Tablets with totals	Discrepant	Rate
Susa	266	19	7.1%
Anshan	41	4	9.8%
Sialk	18	2	11.1%
Other/peripheral	15	2	13.3%

The mildly elevated discrepancy rate at peripheral sites is consistent with, though far from conclusive evidence for, a hypothesis of somewhat less rigorous or less centrally standardized accounting practice away from the Susa administrative core, a theme that recurs in the network analysis of Section 1.10.

26.3 A Probabilistic Model of Sign-Substitution Error

Beyond arithmetic checks, this section models sign-level substitution error using a confusion-matrix approach borrowed from epigraphic error analysis of better-attested scripts. For each pair of graphically similar signs (judged by an independent paleographic similarity coding), the empirical rate at which one form appears in administrative contexts where the other would be expected, based on commodity co-occurrence patterns, is estimated. Of 84 graphically similar sign pairs examined, 11 show a substitution rate elevated significantly above the corpus-wide baseline, consistent with these pairs being genuinely confusable in scribal practice, while the majority show no elevated substitution, supporting the sign list's treatment of these pairs as distinct rather than as undetected allographs.

26.4 Implications for the Slot Grammar and Hypothesis Evaluation

The relatively low and site-comparable error rate documented in this chapter provides modest further support for the H2 (constrained administrative semasiography) hypothesis evaluated in Chapter 5: a system with widespread, poorly controlled scribal error would be harder to reconcile with the high degree of structural regularity (slot adherence, metrological consistency) documented elsewhere in this monograph.

© Silvano Salvador

Chapter 27

A Quantitative Theory of Sign Obsolescence and Innovation

27.1 Sign Turnover Across the Three-Phase Chronology

This chapter examines the rate at which signs enter and exit the attested Proto-Elamite repertoire across the PE I–III chronological sequence developed in Section 1.7, applying methods adapted from lexical-turnover studies in historical linguistics.

Definition: Sign Turnover Rate

For consecutive phases Φ_i and Φ_{i+1}, the *turnover rate* is $\tau_i = \frac{|S_i \setminus S_{i+1}| + |S_{i+1} \setminus S_i|}{|S_i \cup S_{i+1}|}$, where S_i is the set of signs attested at least twice in phase Φ_i.

Table 27.1: Sign turnover rates between adjacent chronological phases

Transition	$\lvert S_i \setminus S_{i+1} \rvert$	$\lvert S_{i+1} \setminus S_i \rvert$	τ
PE I $\rightarrow$ PE II	38	52	0.21
PE II $\rightarrow$ PE III	29	41	0.16

The somewhat lower turnover between PE II and PE III is consistent with, though not conclusive proof of, a maturing and stabilizing notational system over the attested chronological span, paralleling the slightly declining diachronic entropy trend noted (with appropriate statistical caution) in Section 2.6.

27.2 Newly Innovated Signs and Their Functional Profile

Examining the 52 signs newly attested in PE II relative to PE I, a disproportionate share (61%, versus 38% in the corpus generally) fall into Class III/IV (rare, locally restricted signs per Section 12.4) rather than becoming high-frequency Class I or II signs, suggesting that most notational innovation across this transition reflects elaboration of specialized local or commodity-specific vocabulary rather than restructuring of the system's core grammar, consistent with the slot grammar's core slots remaining stable in their occupant-sign sets across phases (Section 14.6).

27.3 Obsolescence and the Limits of the Available Evidence

Signs that disappear from the attested record between phases cannot be securely distinguished from signs that remain in use but are simply absent from the necessarily incomplete excavated and published sample for a given phase. The turnover rates reported above should therefore be read as upper bounds on true obsolescence, inflated by an unknown but plausibly substantial sampling component; Section 23's capture–recapture analysis is relevant context for calibrating the likely scale of this effect.

© Silvano Salvador

Chapter 28

The Economics of Scribal Labor: Workload, Specialization, and Output

28.1 Estimating Scribal Hand Count and Productivity

Building on the scribal-hand variation analysis of Section 2.5, this chapter estimates the implied labor economics of Proto-Elamite administrative record-keeping. The paleographic clustering identified there yields an estimate of 34–41 distinct scribal hands represented in the Susa corpus across the full chronological span, with individual hands contributing between 3 and 187 tablets each — a heavily right-skewed distribution consistent with a professional scribal hierarchy in which a small number of highly productive specialists handled the bulk of routine accounting while a larger number of occasional scribes contributed smaller volumes.

Table 28.1: Distribution of tablet output across estimated scribal hands, Susa corpus

Output tier	Hands	% of total tablets
High-volume (>50 tablets)	4	41%
Mid-volume (10–50 tablets)	11	38%
Low-volume (<10 tablets)	19–26	21%

28.2 Specialization by Administrative Category

Cross-tabulating estimated scribal hand against administrative document category (Section 13.3) reveals that the four high-volume hands show significantly non-uniform category distributions (chi-square test, $p < 0.01$ for three of four), consistent with at least partial functional specialization — e.g., one high-volume hand accounts for 73% of grain-account tablets but only 12% of personnel accounts in the sample, suggesting domain-specific scribal roles within the Susa administrative apparatus broadly analogous to the specialized accounting roles documented ethnographically and textually for the better-attested proto-cuneiform Uruk administration.

28.3 Implications for the Corpus as a Sampling Frame

The concentration of output among a handful of prolific hands has a direct methodological implication for every entropy and frequency statistic computed in this monograph: the corpus

is not a sample of independent scribal voices in proportion to some underlying population of scribes, but is heavily weighted toward the idiolectal habits of a small number of individuals. The robustness program of Chapter 21 addresses this partially through its by-hand subsampling check (an extension of the by-site check reported in Section 16.6, not separately tabulated here), which finds the corpus-wide entropy and Zipf statistics relatively insensitive to exclusion of any single high-volume hand, with no exclusion shifting unigram entropy by more than 0.15 bits.

© Silvano Salvador

Chapter 29

Authenticity, Provenance, and Forgery Detection in the Proto-Elamite Corpus

29.1 The Antiquities-Market Problem

A non-trivial fraction of tablets catalogued as Proto-Elamite entered Western museum and private collections through the antiquities market, with findspot information reconstructed, where possible, from dealer records, stylistic comparison, or, in a small number of cases, not reconstructible at all. This chapter addresses the resulting authenticity and provenance-confidence question directly, since any analysis treating an undetected forgery as genuine evidence risks contaminating every downstream statistic.

Table 29.1: Provenance-confidence classification of the analytical corpus

Provenance tier	Tablets	% of corpus
Excavated, documented findspot	1,198	73.0%
Excavated, findspot reconstructed from records	287	17.5%
Market provenance, site attribution by style	142	8.7%
Market provenance, no secure attribution	13	0.8%

29.2 A Statistical Screen for Anomalous Tablets

Beyond documentary provenance, this section applies a purely internal statistical screen: for each tablet, a likelihood score is computed under the corpus-wide n-gram model of Chapter 3, and tablets whose score falls more than three standard deviations below the corpus mean (after controlling for tablet length and administrative category) are flagged for closer paleographic review. Of the 13 tablets with no secure attribution in Table 21.1, 2 are flagged by this screen; both are excluded from the headline analyses of Chapters 2 through 6 as a conservative precaution, though their exclusion does not measurably affect any reported statistic given the small number involved.

Scope of the Screen

This statistical screen is a coarse triage tool, not a forgery-detection method in any strong sense: a competent forgery constructed from genuine sign exemplars, in genuine combinatorial patterns, would not necessarily produce an anomalous likelihood

score. The screen is best understood as flagging tablets whose sign sequences are unusual relative to corpus norms for any reason, forgery being only one possible explanation among several (genuine but atypical administrative content, transcription error, or an unusual document genre).

29.3 Sensitivity of Headline Results to the Market-Provenance Subcorpus

Recomputing the unigram entropy and Zipf-exponent statistics excluding all 155 market-provenance tablets (Table 21.1, rows 3–4) produces estimates within the bootstrap confidence intervals reported in Table 16.2, indicating that the headline statistical findings of this monograph do not depend materially on the inclusion of tablets lacking secure excavation provenance.

© Silvano Salvador

Chapter 30

A Comparative Survey of Decipherment Methodologies Across Scripts

30.1 Three Historical Pathways to Decipherment

Surveying the small number of ancient scripts successfully deciphered since the nineteenth century reveals three broad methodological pathways, each instructive for assessing Proto-Elamite's prospects.

Table 30.1: Pathways to decipherment among comparandum scripts

Pathway	Example	Key enabling evidence
Bilingual/multilingual text	Egyptian hieroglyphs (Rosetta Stone); Linear B (proper-name matching against known Greek)	A parallel text in a known language or script
Phonetic bootstrapping from a known successor language	Linear Elamite (via Old Persian/Elamite cuneiform correspondences)	A historically attested descendant or closely related language
Internal structural decipherment without external bilingual	Proto-cuneiform numerical systems (Friberg); arguably Maya partial decipherment stages	Sufficient internal redundancy and arithmetic/structural constraint to triangulate meaning

30.2 Why Proto-Elamite Has Access to None of the Three Pathways at Scale

Proto-Elamite possesses no known bilingual text of any kind; its successor relationship to Linear Elamite, while plausible on geographic and chronological grounds, has not been demonstrated with the philological rigor that connected Old Persian to Elamite cuneiform; and its main sign system, unlike its numerical sub-system, lacks the internal arithmetic redundancy that permitted Friberg's success. This survey provides external, comparative corroboration

for the more pessimistic strand of the hypothesis evaluation in Chapter 5 and the provisional posterior ranking of Chapter 6, which assign substantial combined weight to H2 and H4 — hypotheses compatible with permanent or near-permanent non-decipherment of the main sign system's specific semantic content, even while its broad structural properties continue to be illuminated by quantitative methods.

30.3 What Would Change the Picture

Three categories of future discovery would materially alter this assessment: (i) a genuine bilingual or trilingual text pairing Proto-Elamite with a known script, none of which has been found despite over a century of excavation; (ii) a substantially larger and more diverse corpus permitting much higher statistical power for the n-gram and Bayesian methods of Chapters 3 and 6; and (iii) a secure, philologically rigorous demonstration of genetic continuity between Proto-Elamite and a later, better-understood language, which would permit phonetic bootstrapping analogous to the Linear Elamite case. None of these is foreclosed by present evidence, but none is currently in hand either.

© Silvano Salvador

Chapter 31

Quantitative Iconography of the Glyptic and Sealing Corpus

31.1 Seal Impressions as a Parallel Information Channel

Independent of the textual sign system, Proto-Elamite administrative tablets frequently bear seal impressions, and the glyptic record more broadly supplies a second, non-linguistic channel of administrative information already exploited for chronological purposes in Section 1.6. This chapter extends that treatment with a quantitative analysis of seal-design motif frequency and co-occurrence with textual content.

Table 31.1: Principal glyptic motif categories and tablet co-occurrence (n = 380 sealed tablets in sample)

Motif category	Frequency	Co-occurs predominantly with
Geometric/abstract patterning	41%	Grain accounts
Animal-file processions	28%	Animal accounts
Combat/contest scenes	19%	Mixed administrative
Figural/anthropomorphic	12%	No strong category association

31.2 Statistical Association Between Motif Category and Administrative Content

A chi-square test of independence between motif category and the administrative document category of the sealed tablet (categories per Section 13.3) rejects independence at $p < 0.001$, with geometric/abstract motifs strongly overrepresented on grain accounts and animal-file motifs strongly overrepresented on animal accounts. This association suggests that seal choice was not purely a matter of personal or institutional identity unrelated to document content, but tracked, at least loosely, the administrative domain of the transaction being recorded — a finding with implications for the identifier-sign hypothesis (H3) of Section 5.3, since it indicates that non-textual administrative signaling of document category was already present in the Proto-Elamite system through a channel entirely independent of the sign sequence itself.

31.3 Implications for the Slot Grammar and H3

If seal motif already encodes administrative category information redundantly with whatever the text itself encodes, this weakens (though does not eliminate) the case for interpreting any specific textual sign as a dedicated category-identifier under H3, since the same informational function could be served by the glyptic channel rather than by a hypothesized textual identifier sign. This consideration is incorporated, as a downward adjustment to the H3 prior, in the sensitivity analysis of Section 6.5.

© Silvano Salvador

Chapter 32

Typological Universals and Proto-Elamite: A Cross-System Test

32.1 Candidate Universals From the Typology of Early Administrative Notation

Drawing on the comparative typology developed in Chapter 17 and on a broader cross-script literature (Houston's work on inventions of writing; Baines's comparative survey of early notational systems), this chapter tests whether Proto-Elamite conforms to several proposed cross-cultural universals of early administrative writing.

Table 32.1: Proposed universals of early administrative notation and Proto-Elamite's conformity

Proposed universal	Conforms?	Note
Numerical notation precedes or co-emerges with general notation	Yes	Numerical signs attested from earliest phase
Earliest texts are administrative/economic, not narrative	Yes	No narrative content identified in any phase
Vocabulary size grows then stabilizes across early phases	Partial	Stabilization evident PE II→III; earlier growth less well constrained
Sign-frequency distribution approximates Zipfian form	Partial	Near-Zipfian with flattened tail (Section 2.2)
Positional/syntactic regularity increases with administrative complexity	Yes	Slot grammar coverage highest at Susa, lowest at peripheral sites

32.2 Interpretation: A System Typologically Unremarkable in Its Origins, Distinctive in Its Trajectory

The pattern in Table 24.1 situates Proto-Elamite as broadly typologically unremarkable in the conditions of its emergence — consistent with cross-script generalizations about how administrative notation systems originate — while its subsequent trajectory, characterized by several centuries of apparent stability followed by replacement rather than gradual evolution

into a more fully linguistic system (the fate of cuneiform), remains comparatively distinctive and is one of the more striking unresolved features of its history, discussed further in the synthesis of Chapter 10.

© Silvano Salvador

Chapter 33

Pedagogical Appendix: Worked Statistical Derivations for Non-Specialists

33.1 Purpose of This Appendix

The chapters of this monograph assume familiarity with information theory, Bayesian model comparison, and basic time-series statistics. This appendix is addressed to readers trained primarily in philology or archaeology who wish to follow the quantitative argument in detail without first consulting a separate statistics textbook. It works through, step by step, the derivations underlying the four most heavily used formal tools in the monograph.

33.2 Deriving Shannon Entropy From First Principles

Worked Derivation: Why $H = -\sum p_i \log_2 p_i$

Suppose a sign sequence is generated by independently drawing each sign from a fixed distribution $p_1, \dots, p_V$ over a vocabulary of size V. The number of distinct sequences of length n that are "typical" — consistent with the expected frequency of each sign — grows, for large n, as 2^{nH} where $H = -\sum_i p_i \log_2 p_i$. This follows from Stirling's approximation applied to the multinomial coefficient counting sequences with sign i appearing np_i times: $\binom{n}{np_1, \dots, np_V} \approx 2^{nH}$. The entropy H therefore measures, in bits, the average number of yes/no questions needed to identify which sign occupies a given position, given only the marginal frequency distribution.

33.3 Deriving the Bayes Factor From the Likelihood Ratio

Worked Derivation: Bayes Factor as Evidence Update

Given two hypotheses H_a and H_b and data D, Bayes' theorem gives posterior odds as $\frac{P(H_a|D)}{P(H_b|D)} = \frac{P(D|H_a)}{P(D|H_b)} \times \frac{P(H_a)}{P(H_b)}$. The first factor on the right, the Bayes factor $K = P(D \mid H_a)/P(D \mid H_b)$, is exactly the factor by which the data update the prior odds into posterior odds; it depends only on the data's relative likelihood under the two hypotheses, not on the prior itself, which is why Chapter 6 reports Bayes factors separately from

the prior-dependent posterior probabilities, allowing readers who disagree with the elicited priors of Section 6.1 to substitute their own and recompute the posterior directly from the reported K values.

33.4 Deriving the Minimum Description Length Criterion

Worked Derivation: MDL as Compressed Code Length

For any probability model P assigning likelihood $P(D)$ to data D, Shannon's source coding theorem guarantees the existence of a code assigning D a codeword of length $-\log_2 P(D)$ bits (the Shannon–Fano construction), and no code using P as its probability assignment can do appreciably better on average. The MDL criterion of Section 6.3 compares two models by their total description length — the cost of encoding the model's parameters plus the cost of encoding the data given the model — and prefers the model minimizing this sum, formalizing an Occam's-razor intuition: a model that fits the data well but requires many free parameters to do so may lose to a simpler model that fits slightly less well, once the cost of specifying the extra parameters is counted explicitly.

33.5 Deriving the Permutation-Test p-value

Worked Derivation: Why Permutation Tests Need No Distributional Assumption

To test whether an observed statistic $\hat{\theta}$ (e.g., a cross-entropy gap between two sub-corpora) is larger than chance would produce, the permutation test of Section 15.3 randomly reassigns the sub-corpus labels many times, recomputing $\hat{\theta}$ under each random relabeling, and reports the fraction of relabelings producing a statistic as extreme as the one actually observed. Because the relabeling is performed on the actual data rather than on a theoretical reference distribution, no assumption about the underlying sampling distribution (normality, for instance) is required; the validity of the resulting p-value rests only on the assumption that, under the null hypothesis of no true difference, all relabelings are equally probable.

© Silvano Salvador

Chapter 34

Computational Stemmatics: Scribal Schools and Sign-Form Transmission

34.1 Adapting Stemmatic Methods From Manuscript Textual Criticism

This chapter applies methods originally developed for reconstructing manuscript transmission trees in textual criticism — a field concerned with inferring the copying relationships among surviving manuscript witnesses from their patterns of shared variant readings — to the problem of reconstructing relationships among Proto-Elamite scribal hands based on shared idiosyncrasies in sign-form execution and slot-filling convention.

Definition: Scribal Distance

For two scribal hands h_1, h_2 identified by the paleographic clustering of Section 2.5, the *scribal distance* $d(h_1, h_2)$ is the proportion of a fixed battery of diagnostic features — sign-form allograph choice, slot-filling order preference, and metrological-system preference among interchangeable options — on which the two hands disagree.

34.2 A Neighbor-Joining Tree of Scribal Hands

Applying neighbor-joining clustering to the pairwise scribal-distance matrix computed over the 34–41 estimated hands of Chapter 28 produces a branching structure suggestive of two or three loosely defined "schools," the largest of which accounts for roughly half of all hands and is associated predominantly, though not exclusively, with grain-account specialization.

Methodological Caveat

Unlike textual stemmatics, where shared error against an inferable original reading provides strong evidence of direct copying relationships, scribal-hand similarity in an administrative notational practice may equally well reflect shared training under a common institutional convention, geographic proximity, or simple chronological contemporaneity, rather than any direct teacher-pupil relationship. The neighbor-joining tree should be read as a description of similarity structure, not as a claim about literal transmission lineages.

Chapter 35

Climate, Destruction Horizons, and the End of the Proto-Elamite Tradition

35.1 The PE III to Linear Elamite Transition: An Abrupt Discontinuity

The Proto-Elamite notational tradition, as quantitatively characterized throughout this monograph, does not shade gradually into its eventual successor system, Linear Elamite (Chapter 7), but rather ends, at Susa and elsewhere, in a horizon associated archaeologically with significant cultural and political discontinuity at the close of the PE III phase, followed by a multi-century gap before Linear Elamite's appearance in the much later third-millennium record.

35.2 Environmental and Political Context

Paleoclimatic proxy evidence from the broader southwest Iranian and Mesopotamian region indicates a period of increased aridity in the later third millennium associated, in neighboring Mesopotamia, with documented agricultural and settlement stress. Whether comparable climatic pressure contributed to the administrative system's discontinuation at Susa, as opposed to political factors connected to the rise of competing polities in the Iranian plateau and lowland Mesopotamia, cannot be resolved on the textual evidence examined in this monograph and is flagged here as an open question for archaeological rather than computational-linguistic methods to address.

35.3 Implications for the Corpus as a Closed, Bounded Dataset

For the purposes of the quantitative methods employed throughout this monograph, the principal significance of the discontinuous end to the Proto-Elamite tradition is methodological rather than historical: the corpus analyzed here is a closed, temporally bounded dataset, not a sample drawn from an open-ended, continuously evolving system. This justifies treating phase PE III as a genuine endpoint for the diachronic trend analyses of Section 2.6 and Chapter 27, rather than as an arbitrary truncation of an ongoing process.

Chapter 36

A Formal Theory of Scribal Memory: Recurrence and Long-Range Dependence

36.1 Beyond Local n-gram Structure: Testing for Long-Range Dependence

The n-gram models of Chapter 3 capture only local, fixed-window sequential dependence. This chapter tests whether Proto-Elamite sign sequences exhibit long-range dependence beyond what any finite-order Markov model can capture, using a Hurst-exponent estimation applied to the binary indicator sequence of a given sign's occurrence across a tablet's sign positions.

Table 36.1: Estimated Hurst exponents for selected high-frequency signs

Sign (catalogue ID)	Hurst exponent $\hat{H}$	95% CI
M388 (grain commodity marker)	0.61	[0.54, 0.68]
M001 (general counting sign)	0.52	[0.46, 0.58]
M295 (candidate identifier, Section 5.3)	0.58	[0.49, 0.67]

A Hurst exponent significantly above 0.5 indicates persistent, long-range positive autocorrelation in a sign's occurrence pattern, beyond what local Markov structure would predict. The result for M388, with a confidence interval excluding 0.5, is consistent with this commodity sign clustering at a longer range than slot-grammar adjacency alone would explain, plausibly reflecting tablet-level thematic coherence (a tablet about grain tends to mention grain-related signs repeatedly throughout, not just in adjacent slots).

36.2 Implications for the Cross-Entropy and Transfer-Entropy Analyses

The presence of long-range dependence for at least some signs implies that the fixed-order n-gram and transfer-entropy estimates of Chapters 3–4 are likely conservative (i.e., likely to understate) the true predictability of the sign sequence, since these methods are by construction blind to dependence beyond their fitted order. This does not overturn any qualitative conclusion of those chapters but suggests that the true cross-entropy gap against the proto-

cuneiform reference corpus (Section 3.2) may be somewhat larger than the fixed-order estimate reported there.

© Silvano Salvador

Chapter 37

Trade Networks and Foreign Contact: Evidence from Outside Susiana

37.1 Proto-Elamite Material Beyond Its Core Distribution

While the great majority of the Proto-Elamite corpus derives from Susa, Anshan, and Sialk (Section 1.10), isolated finds of Proto-Elamite tablets and, more frequently, Proto-Elamite-style sealings have been reported from sites considerably further afield across the Iranian plateau, raising questions about the geographic extent of the administrative and exchange network within which the script operated.

Table 37.1: Reported Proto-Elamite or Proto-Elamite-influenced finds outside the core distribution

Site	Distance from Susa	Nature of evidence
Tepe Yahya	∼600 km SE	Sealings with Proto-Elamite-style motifs; no tablets
Shahr-i Sokhta	∼900 km E	Indirect glyptic parallels; contested attribution
Konar Sandal	∼650 km SE	Sealing fragments, debated dating

37.2 What the Peripheral Glyptic Evidence Does and Does Not Establish

The presence of Proto-Elamite-style sealing practice at considerable distances from the Susa core is compatible with several distinct historical scenarios: direct administrative integration of these sites into a Susa-centered network (unlikely given the distances and the absence of any textual tablets at the furthest sites); emulation of a prestigious neighboring administrative technology without integration into its actual record-keeping apparatus; or exchange-mediated transmission of sealing practice through intermediary trade partners without either side adopting the other's full notational system. The quantitative methods of this monograph, dependent on textual tablet content, cannot adjudicate among these scenarios and this chapter's contribution is limited to cataloguing the relevant evidence and its geographic extent.

37.3 Implications for the Network Analysis of Section 1.10

The network analysis of administrative flows in Section 1.10 was restricted, by necessity, to sites with tablet-bearing assemblages, and therefore likely understates the true geographic reach of Proto-Elamite administrative and glyptic influence. The peripheral sealing evidence catalogued here suggests that this influence extended, in some attenuated form, considerably beyond the network graph's documented node set.

© Silvano Salvador

Chapter 38

Material Analysis of Inscribed Surfaces and Its Evidentiary Value

38.1 Clay Fabric Analysis Revisited

Section 1.8 introduced tablet fabric and stylus-impression morphometrics as a chronological tool. This chapter extends that material-science treatment to ask what fabric composition can reveal about administrative organization independent of dating.

38.2 Clay Sourcing and Administrative Locality

Petrographic and, for a subset of the corpus, portable XRF compositional analysis of tablet clay fabric at Susa identifies at least two statistically distinguishable clay sources in use across the sampled tablets, with the relative proportion of each source shifting across the three chronological phases in a pattern broadly consistent with, though not strictly required by, the independent ceramic-fabric chronology of Section 1.8.

Limits of the Compositional Evidence

Clay source does not straightforwardly indicate scribal identity or administrative unit, since a single source could supply multiple administrative offices and a single office could draw on multiple sources over time. The compositional data are presented here as a further, independent line of chronological corroboration rather than as evidence bearing on administrative organization per se.

38.3 Stylus Wear and the Reuse of Writing Implements

Detailed measurement of impression-edge wear across large single-hand tablet groups (the high-volume scribal hands of Chapter 28) shows progressive blunting consistent with sustained reuse of a small number of styli per scribe over an extended working period, supporting the characterization of the high-volume hands as professional specialists working over a substantial career span rather than as an aggregation of distinct short-term scribes whose work has been mis-clustered by the paleographic method of Section 2.5.

Chapter 39

Toponymic and Institutional Evidence: Place and Office Designations

39.1 Candidate Toponymic and Institutional Signs

A small number of signs recur across multiple sites with a distribution pattern suggestive of place-name or institutional-office reference rather than commodity or quantity content: appearing at the periphery of the slot grammar (Section 14.6, outside the five core administrative slots), with low overall frequency but high site-specificity.

Table 39.1: Candidate toponymic/institutional signs and their site distribution

Sign (catalogue ID)	Susa freq.	Anshan freq.	Site-specificity index
M512	0.4%	8.1%	0.89
M647	0.2%	0.1%	0.91 (Sialk-specific)
M203	1.1%	0.3%	0.62

39.2 Caution Regarding Onomastic Interpretation

The strong site-specificity of signs such as M512 and M647 is consistent with, but does not by itself establish, a toponymic or institutional reading: an alternative explanation is that these signs denote locally distinctive commodities or administrative categories specific to each site's economic profile rather than the site's name or a local office per se. Absent any independent check comparable to Friberg's arithmetic anchor for the numerical system (Chapter 22), this monograph treats the toponymic reading as a live but unconfirmed hypothesis, relevant chiefly to H3 (Section 5.3) and incorporated into that hypothesis's evidentiary discussion rather than asserted as established fact.

Chapter 40

Quantitative Cross-Script Benchmarking: A Wider World Sample

40.1 Extending the Reference Corpus Beyond Proto-Cuneiform and the Synthetic Control

Chapter 7 compares Proto-Elamite against proto-cuneiform and a synthetic semasiographic control. This chapter widens the comparison to a broader sample of early and undeciphered notational systems, computing the same entropy and Zipf statistics uniformly across all systems where adequate digitized corpora exist.

Table 40.1: Comparative entropy and Zipf statistics across a wider script sample

System	Unigram entropy (bits)	Zipf exponent	Corpus size (tokens
Proto-Elamite (this study)	7.92	1.07	17,400
Proto-cuneiform (reference)	7.35	1.18	154,000
Indus Valley script	7.61	1.04	11,200
Linear A (companion volume)	8.05	1.02	21,800
Rongorongo (illustrative, low confidence)	7.18	0.95	4,900

40.2 Positioning Proto-Elamite Within the Wider Sample

Proto-Elamite's entropy and Zipf statistics fall within the same broad range as the other early administrative and undeciphered systems surveyed, situating it as unremarkable on these particular dimensions relative to its closest typological peers, and reinforcing the classification-space analysis of Chapter 5 (Section 12.3 onward): whatever distinguishes Proto-Elamite's specific decipherment difficulty from, say, proto-cuneiform's eventual decipherment is unlikely to be visible in corpus-level entropy or Zipf statistics alone, and more likely resides in the specific absence of external anchoring evidence (Chapter 30) rather than in any distinctive internal statistical signature.

Chapter 41

A Formal Model of the Proto-Elamite Administrative Cycle

41.1 Seasonal and Periodic Structure in Administrative Content

Agricultural administration in the ancient Near East typically follows a seasonal accounting cycle tied to harvest and redistribution. This chapter examines whether the Proto-Elamite corpus exhibits internal evidence of such periodicity, despite the absence of any explicit calendrical notation comparable to later cuneiform date formulas.

41.2 Indirect Evidence: Commodity Co-occurrence Clustering as a Cycle Proxy

Lacking explicit dates, this analysis uses commodity-sign co-occurrence clustering (extending the metrological-system partition of Section 17.3) as an indirect proxy: tablets dominated by grain-harvest-adjacent commodity signs are hypothesized to cluster temporally around a harvest accounting period, distinct from tablets dominated by animal-husbandry signs, which need not follow the same periodicity.

Strength of Inference

This indirect approach can establish, at most, that the corpus contains thematically distinct administrative sub-episodes; it cannot establish their absolute or even relative timing within a calendar year, since no Proto-Elamite tablet carries an explicit seasonal or calendrical marker recognized as such by current scholarship. The chapter's contribution is accordingly limited and clearly flagged as such.

41.3 A Negative but Informative Result

Clustering analysis of commodity co-occurrence does identify the grain/discrete-goods partition already documented in Section 17.3, but does not reveal any further, finer-grained periodic substructure within the grain-account cluster that would support a specific seasonal-cycle reconstruction. This negative result is informative: it suggests that whatever administrative periodicity existed in practice left no statistically detectable trace in the surviving sign-sequence record at the resolution available to this corpus, distinguishing Proto-Elamite sharply from later cuneiform administrative archives where explicit month names make seasonal reconstruction straightforward.

Chapter 42

Diachronic Entropy Drift Across the Proto-Elamite Sequence

42.1 Revisiting the Diachronic Trend With the Full Toolkit

Section 2.6 introduced a preliminary diachronic entropy comparison across the three chronological phases. This chapter revisits that comparison using the robustness and imputation machinery developed in Chapters 21–23, to assess whether the mild downward entropy trend reported there survives more careful treatment of phase-assignment uncertainty and missing data.

Table 42.1: Diachronic entropy estimates with combined phase-assignment and missing-data uncertainty

Phase	Point estimate (bits)	Combined 95% CI	N (tablets)
PE I	8.04	[7.78, 8.31]	198
PE II	7.89	[7.74, 8.05]	891
PE III	7.81	[7.61, 8.02]	524

The PE I to PE III decline remains directionally consistent after accounting for the additional uncertainty sources, though the PE I estimate's wide interval (reflecting its much smaller documented sample) means the trend cannot be considered firmly established at conventional significance thresholds when PE I is included; restricting to the PE II/PE III comparison alone, where sample sizes are larger, yields a more confidently estimated, modest decline.

42.2 A Candidate Explanation: Administrative Standardization Over Time

The most parsimonious explanation consistent with the available evidence is gradual administrative standardization: as the slot-grammar conventions documented in Section 14.6 became more entrenched in scribal practice (supported also by the declining sign-turnover rate of Chapter 27), the effective predictability of sign choice at a given position would be expected to rise, lowering measured entropy, without requiring any change in the underlying semantic content being recorded.

Chapter 43

Coda: What This Monograph Can and Cannot Establish

43.1 A Summary Ledger

In closing, it is useful to state plainly, in one place, the epistemic status of the principal claims advanced across the preceding chapters.

Table 43.1: Epistemic status of principal claims

Claim	Status	Basis
Proto-Elamite numerical notation is substantively deciphered	Established	Friberg's arithmetic-anchored reconstruction
The main sign system encodes a specific natural language	Undetermined	No statistical method in this monograph can adjudicate this
A five-slot administrative grammar describes the bulk of transactional tablets	Well-supported	Section 14.6, with 78% coverage
A restricted-register positional anomaly exists and is not a corpus-size artifact	Well-supported	Section 2.3; Chapter 21
H2/H4 jointly dominate the posterior over H1/H3	Provisional	Sensitive to prior elicitation (Section 6.5)
The script will likely never be deciphered absent new bilingual evidence	A reasoned judgment, not a proof	Chapter 30

43.2 On the Value of Quantitative Method in the Absence of Decipherment

A reader who approaches this monograph hoping for a decipherment, or even a confident prediction of imminent decipherment, will be disappointed, and that disappointment is, in the authors's judgment, the intellectually honest outcome of applying rigorous quantitative method to the available evidence rather than a failure of the method itself. What the quantitative program developed across these chapters does deliver is a calibrated, falsifiable, and re-

producible characterization of what the surviving Proto-Elamite corpus actually contains: its information-theoretic structure, its administrative grammar, its chronological development, and the specific, well-defined ways in which it resembles and differs from both its deciphered and its undeciphered typological neighbors. This is a more modest contribution than a decipherment, but it is one that future decipherment attempts, should new evidence ever permit them, will need to be consistent with, and one that, in its absence, still substantially advances understanding of one of the world's most durable archaeological puzzles.

© Silvano Salvador

Chapter 44

Additional Case Studies: Further Close Readings

44.1 Case Study 13: A Multi-Commodity Susa Tablet and Slot-Grammar Saturation

CDLI P008912, a well-preserved PE II Susa tablet recording a multi-line grain and labor account, instantiates all five slots of the administrative grammar (Section 14.6) on every one of its eleven recorded line entries, making it one of the cleanest available illustrations of full slot saturation. The tablet's stated total matches the sum of its line entries exactly, providing a useful positive control against which the arithmetic-error detection method of Section 26 can be calibrated.

44.2 Case Study 14: An Anshan Tablet Exhibiting the Discrete-Goods Metrological Cluster

CDLI P010512 records a personnel account using System S (Table 17.1) throughout, with commodity signs drawn exclusively from the discrete-goods cluster identified in Section 17.3. Its slot-grammar coverage (Slots 1, 2, 4, and 5 present; Slot 3 absent on six of eight lines) is consistent with the lower average coverage rate documented for peripheral sites in Section 14.6, and offers a concrete illustration of that aggregate finding.

44.3 Case Study 15: A Tablet Bearing a Candidate Toponymic Sign

CDLI P011033, from Sialk, contains two occurrences of sign M647, the Sialk-specific candidate toponymic sign discussed in Chapter 39. Both occurrences appear at the tablet's margin, outside the five core administrative slots, consistent with the peripheral positional distribution expected of a toponymic or institutional marker rather than a transactional content sign.

44.4 Case Study 16: A Heavily Sealed Tablet and the Iconography–Text Relationship

CDLI P009287 bears three distinct seal impressions alongside a grain account written in a single identifiable scribal hand. All three seals fall into the geometric/abstract motif category of Table 22.1, consistent with the documented association between that motif category and

grain-account content, and providing a tablet-level illustration of the corpus-wide statistical association reported in Chapter 31.

44.5 Synthesis of the Extended Case-Study Sample

Taken together with the twelve case studies of Chapter 9, this extended sample of sixteen close readings illustrates, at the level of individual artifacts, nearly every quantitative pattern documented at corpus scale elsewhere in this monograph: slot-grammar saturation and its site-dependent variation, the discrete-goods/grain metrological partition, candidate toponymic signs in their characteristic peripheral slot position, and the statistical association between glyptic motif and administrative content. No close reading by itself constitutes evidence at the same standard as a corpus-wide statistic, but the convergence between individual-tablet illustration and aggregate pattern across sixteen independently selected examples provides a useful qualitative check on the plausibility of the quantitative findings reported throughout.

© Silvano Salvador

Chapter 45

Extended Data Appendix: Supplementary Tables

45.1 Full Site-by-Phase Tablet Counts

Table 45.1: Tablet counts by site and chronological phase

Site	PE I	PE II	PE III	Undated	Total
Susa	142	712	398	304	1,556
Anshan	31	89	67	22	209
Sialk	18	47	35	9	109
Other/peripheral	7	43	24	12	86
Total	198	891	524	347	1,960

45.2 Full Administrative Document Category Breakdown by Site

Table 45.2: Administrative document category distribution by site (% of site total)

Site	Grain	Animal	Personnel	Mixed	Other
Susa	38%	21%	19%	17%	5%
Anshan	29%	31%	24%	11%	5%
Sialk	33%	28%	18%	14%	7%
Other/peripheral	26%	35%	21%	12%	6%

45.3 Full Metrological System Usage by Commodity Class

45.4 Scribal Hand Output Detail

These supplementary tables consolidate, in a single reference location, disaggregated counts underlying the summary statistics and percentages cited throughout Chapters 1, 12–14, and 17–22, and are provided primarily as a resource for readers wishing to reproduce or extend the corpus-level computations described in Chapter 11.

Table 45.3: Metrological system usage frequency by commodity class

Commodity class	System S	System B	System C	Area system
Grain	4%	86%	6%	4%
Animals	91%	2%	5%	2%
Personnel/labor	88%	3%	7%	2%
Land/area	2%	1%	3%	94%

Table 45.4: Output detail for the four high-volume scribal hands

Hand ID	Tablets	Dominant category	Sites attested	Phase span
H-01	187	Grain (73%)	Susa only	PE II–III
H-02	134	Mixed	Susa only	PE II
H-03	98	Personnel (61%)	Susa, Anshan	PE II–III
H-04	81	Animal (54%)	Susa only	PE I–II

© Silvano Salvador

Chapter 46

Extended Methodological Glossary: Formal Definitions Consolidated

46.1 Purpose

Chapter 14's thematic glossary covers concepts at a discursive level. This chapter consolidates, for quick reference, the formal mathematical definitions of every named quantity introduced across Chapters 20–45, each restated in compact form with a pointer to its first full discussion.

Table 46.1: Consolidated formal definitions, Part I

Term	Formal definition (compact)	First discussed
Perturbation axis	A discrete/continuous analytic choice defining a family of alternative analyses	§16.2
Scribal distance	Proportion of diagnostic features on which two hands disagree	§34
Sign turnover rate	$\tau_i = \frac{\lvert S_i \setminus S_{i+1}\rvert + \lvert S_{i+1} \setminus S_i\rvert}{\lvert S_i \cup S_{i+1}\rvert}$	§27
Held-out decipherment test	Comparison of held-out predictive log-likelihood against an n-gram null	§24
Arithmetic error signal	Mismatch between itemized sum and stated total beyond plausible reading ambiguity	§26

46.2 Cross-Reference Notes

Several quantities defined in this consolidated glossary depend on quantities defined earlier in the monograph's core chapters (notably the entropy estimators of Section 2.1 and the Bayes-factor machinery of Section 6.2); the glossary entries above are intended as compact reminders for a reader who has already encountered the full derivation, not as self-contained introductions, and the pedagogical appendix (Chapter 33) should be consulted first by any reader encountering these quantities for the first time.

Table 46.2: Consolidated formal definitions, Part II

Term	Formal definition (compact)	First discussed
Metrological system	A closed numeral-sign set with fixed conversion ratios used for a commodity class	§22
Multiple imputation (signs)	M-fold sampling of candidate readings for damaged positions, statistic averaged across draws	§23
Hurst exponent (sign occurrence)	Scaling exponent of rescaled-range statistic applied to a sign's binary occurrence sequence	§36
Site-specificity index	Normalized measure of how concentrated a sign's attestation is at a single site	§39

© Silvano Salvador

Chapter 47

Frequently Raised Objections and Responses

47.1 "The Corpus Is Too Small for Reliable Entropy Estimation"

This objection is addressed directly and repeatedly throughout this monograph, most systematically in Chapter 21. The short response: corpus size is a real limitation, reflected throughout in appropriately wide bootstrap confidence intervals (Table 16.2), but it is not a reason to abandon quantitative analysis altogether, since even an imprecisely estimated quantity, honestly reported with its uncertainty, is more informative than no quantitative estimate at all, provided — as this monograph attempts throughout — that the imprecision is never papered over.

47.2 "Bayesian Priors for the Hypotheses in Chapter 6 Are Arbitrary"

The prior elicitation procedure of Section 6.1 is explicitly acknowledged as involving subjective judgment, informed by the historiographical survey of Chapter 20 and the comparative decipherment-methodology survey of Chapter 30. The sensitivity analysis of Section 6.5 directly addresses this objection by reporting how the posterior ranking of hypotheses changes under a range of alternative, defensible prior specifications, and the reader is invited to substitute their own priors into the reported Bayes factors (Chapter 33, Section 25.3) rather than accept the specific priors elicited here as authoritative.

47.3 "Quantitative Methods Cannot Capture What Philologists Already Know"

This monograph does not claim that quantitative methods supersede philological expertise; rather, it claims that quantitative methods can formalize, test, and in some cases extend specific philological observations — the slot grammar of Section 14.6 formalizes observations already informally present in Damerow and Englund's work, for instance — in ways that make those observations falsifiable against a larger evidentiary base than any individual scholar could review tablet-by-tablet. Where quantitative and philological judgment diverge, as in the toponymic-sign question of Chapter 39, this monograph reports the divergence explicitly rather than resolving it by fiat in favor of either source of evidence.

47.4 "The Restricted-Register Anomaly Might Just Reflect Genre Skew"

This is among the more serious objections raised against the central finding of Section 2.3, and is taken up directly in the by-site subsampling check of Section 16.6 and, less directly, in the seasonal-cycle analysis of Chapter 41. The evidence assembled across these chapters supports, but does not conclusively prove, that the anomaly is a structural property of the notational system rather than purely an artifact of the corpus's skew toward grain and animal accounting; a fully decisive resolution of this question would require a substantially larger and more genre-diverse corpus than currently exists.

© Silvano Salvador

Chapter 48

Paleographic Variation and the Geometry of Sign Form

48.1 Beyond the Sign List: Quantifying Allograph Shape Variation

The sign-list editions surveyed in Section 16.2 record discrete categorical decisions about which surface forms count as the same sign. This chapter complements that categorical treatment with a continuous, geometric analysis of sign-form variation, applying outline-shape statistics borrowed from geometric morphometrics — a method already well established in archaeological lithic and ceramic analysis but applied only sparingly to script paleography.

Definition: Elliptic Fourier Shape Descriptor

For a digitized sign-outline trace, an *elliptic Fourier descriptor* of order k represents the outline as a truncated Fourier series in arc-length parameterization, yielding a fixed-length numerical vector that can be compared across sign instances using standard multivariate distance measures regardless of the outline's original orientation or scale.

48.2 A Shape-Distance Clustering of High-Frequency Signs

Applying tenth-order elliptic Fourier descriptors to a sample of 60 instances each of the twelve most frequent non-numerical signs (Class I, Section 12.3) and clustering the resulting shape-distance matrix recovers, for nine of the twelve signs, a single coherent cluster with low within-sign shape variance relative to between-sign variance, supporting the sign list's treatment of these as unitary categories. The remaining three signs show bimodal shape distributions, suggesting that the current sign list may be merging what are, geometrically, two distinguishable allograph populations.

Table 48.1: Shape-variance ratios for sampled high-frequency signs

Sign (catalogue ID)	Within/between variance ratio	Distribution
M001	0.18	Unimodal
M014	0.21	Unimodal
M388	0.52	Bimodal
M090	0.61	Bimodal
M227	0.49	Bimodal
(remaining 7 signs)	0.14–0.26	Unimodal

48.3 Are the Bimodal Signs Site- or Hand-Specific Variants?

Cross-tabulating the two shape modes identified for M388, M090, and M227 against site and scribal-hand assignment (Chapter 28) shows that for M388 and M227, the two shape modes correlate strongly with the Susa/non-Susa site division, consistent with a regional allographic variant rather than free stylistic variation; for M090, the two modes correlate instead with scribal hand, consistent with an idiosyncratic individual writing habit. This distinction matters for the sign-list sensitivity analysis of Section 16.5: a regionally conditioned allograph, if mistakenly split into two signs, would bias site-comparison statistics, while a hand-conditioned allograph would instead bias the scribal-hand clustering of Section 2.5.

48.4 Recommendation for Future Sign-List Revision

On the strength of this geometric evidence, a revised treatment of M090 as two hand-conditioned allographs of a single sign, while retaining M388 and M227 as currently split, is recommended for future sign-list editions, though this monograph retains the current CDLI/DEPCAS sign list throughout its main analyses for consistency with the cited prior literature, flagging this as an area for future refinement rather than retrospectively altering the corpus mid-analysis.

© Silvano Salvador

Chapter 49

Administrative Hierarchy and the Question of Centralized Oversight

49.1 Evidence for and Against a Centralized Susa Administration

A recurring question in Proto-Elamite studies concerns the degree to which the Susa administrative apparatus exercised centralized oversight over the smaller Anshan and Sialk assemblages, as opposed to each site maintaining a largely autonomous local accounting tradition loosely connected by shared notational convention. This chapter synthesizes evidence bearing on this question already developed piecemeal in earlier chapters — the network analysis of Section 1.10, the scribal-hand site distribution of Chapter 28, and the slot-grammar coverage gradient of Section 14.6 — into a single explicit assessment.

Table 49.1: Evidence bearing on centralized versus autonomous administration

Evidence type	Pattern observed	Favors
Network flow analysis (§1.10)	Asymmetric flow concentrated toward Susa	Centralized
Scribal hand site distribution (Ch. 28)	No high-volume hand attested outside Susa	Centralized
Slot-grammar coverage gradient (§14.6)	Lower coverage at periphery	Mixed/autonomous
Metrological system usage (§17.3)	Consistent across all sites	Centralized (shared standard)
Arithmetic error rate by site (§26)	Elevated at periphery	Mixed/autonomous

49.2 A Synthesis: Standardized Convention, Locally Variable Execution

The pattern across Table 28.1 supports neither a purely centralized model (in which peripheral sites would be expected to match Susa coverage and error rates closely) nor a purely autonomous model (in which the consistent metrological-system usage and absence of locally distinct numerical systems would be harder to explain). The evidence instead favors an intermediate model: peripheral sites operated within a shared notational and metrological

standard transmitted from, or in coordination with, the Susa administrative core, while exercising a degree of local autonomy in day-to-day execution that produced somewhat lower slot-grammar adherence and somewhat higher arithmetic error rates than the Susa core itself.

49.3 Comparison to the Proto-Cuneiform Uruk-Periphery Relationship

A structurally similar pattern — standardized convention transmitted outward from a primary center, combined with locally variable execution at peripheral sites — has been documented for the proto-cuneiform Uruk administrative network by Englund and Nissen, suggesting that this particular center-periphery dynamic may reflect a general feature of early Mesopotamian and Iranian administrative organization in the late fourth and early third millennia rather than anything Proto-Elamite-specific.

© Silvano Salvador

Chapter 50

Personnel Accounts and the Demography of Labor Notation

50.1 The Personnel-Account Sub-Corpus

Personnel and labor accounts constitute roughly one-fifth of the corpus (Table 20.2) and merit dedicated treatment given their potential bearing on questions of Proto-Elamite social and economic organization that lie beyond the script's notational structure narrowly construed.

50.2 Quantitative Profile of Personnel-Account Numerals

Personnel accounts employ System S (Table 17.1) overwhelmingly, with recorded group sizes ranging from single individuals to groups exceeding 100, the latter occurring predominantly in Susa tablets associated with the high-volume scribal hand H-03 (Table 26.4). The distribution of recorded group sizes is strongly right-skewed, with a median of 6 and a small number of large outlier entries plausibly representing institutional labor levies rather than individual household units.

Table 50.1: Distribution of recorded personnel-account group sizes

Group size range	Frequency	% of personnel tablets
1–5	218	51%
6–20	142	33%
21–100	58	14%
>100	9	2%

50.3 No Evidence for Gendered or Age-Differentiated Notation

A natural comparandum is the proto-cuneiform and later cuneiform practice of notating labor-account entries by age and sex category using dedicated sign combinations. Systematic search for analogous sign-combination patterning in the Proto-Elamite personnel-account sub-corpus finds no clear candidate: no recurring sign pair or sign-plus-numeral-modifier pattern shows the kind of restricted, category-like distribution that would support a gendered or age-differentiated reading, though the absence of evidence here cannot be treated as strong evidence of absence given the corpus's limited size and the inherent difficulty of detecting a rare, low-frequency notational distinction statistically.

Caution Against Overinterpretation

Demographic notation, where it exists in comparable early administrative systems, is typically conveyed by a small, dedicated, low-frequency sign subset rather than by the high-frequency signs whose distributional behavior this monograph's methods are best powered to detect. A definitive answer to this question may require targeted philological search rather than the corpus-wide statistical methods used throughout.

© Silvano Salvador

Chapter 51

Comparative Weight and Capacity Systems: Proto-Elamite Against Its Near Eastern Neighbors

51.1 Placing the Proto-Elamite Metrological Systems in Regional Context

Chapter 22 reconstructed the internal structure of Proto-Elamite numerical notation. This chapter situates that reconstruction within the broader landscape of contemporary and near-contemporary Near Eastern weight and capacity systems, asking whether Proto-Elamite numeracy shows evidence of direct structural borrowing from, or independent parallel development relative to, its proto-cuneiform neighbor.

Table 51.1: Base-structure comparison of capacity systems

System	Base structure	System
Proto-Elamite System B (grain)	1–6–60	This study, Ch. 22
Proto-cuneiform capacity system	1–10–... (multiple variants)	Nissen, Damerow, Englund (1993)
Proto-Elamite System S (discrete)	1–10–60–600	This study
Proto-cuneiform sexagesimal	1–10–60–600–3600	Established literature

51.2 Partial Structural Convergence Without Identity

The Proto-Elamite discrete-counting System S shares its base-60 upper structure with the well-attested proto-cuneiform sexagesimal system, a convergence unlikely to be coincidental given the geographic proximity and broadly contemporary dating of the two traditions, and consistent with at least some degree of shared numerical culture across the Susa-Uruk interaction sphere documented archaeologically through exchanged material culture (Section 16). At the same time, the grain-specific System B's distinctive base-6 intermediate step has no precise proto-cuneiform counterpart among the systems catalogued by Nissen, Damerow, and Englund, suggesting an independently developed, Susa-specific capacity convention layered onto a more broadly shared base-60 counting tradition.

51.3 Implications for the Question of Script Origin

This partial convergence is relevant to, though it does not resolve, the broader question (raised in Section 8 and discussed further in Chapter 32) of whether Proto-Elamite notation as a whole developed independently of proto-cuneiform or under its influence. The numerical evidence assembled here supports a middle position: partial, plausibly contact-mediated convergence in basic counting structure, combined with clearly independent development of commodity-specific capacity conventions, a pattern broadly consistent with the main text's general characterization (Chapter 10) of Proto-Elamite as a system developed in a region of active exchange with, but not administrative subordination to, contemporary Uruk.

© Silvano Salvador

Chapter 52

An Agent-Based Simulation of Scribal Practice

52.1 Motivation: From Statistical Description to Generative Mechanism

The preceding chapters characterize the Proto-Elamite corpus's statistical properties in detail but largely stop short of proposing an explicit generative mechanism that could produce those properties. This chapter constructs a minimal agent-based simulation of scribal practice, in the spirit of generative models used in computational sociolinguistics, to test whether a simple set of behavioral rules can jointly reproduce several of the corpus's headline statistical signatures without being fitted directly to any of them.

52.2 Model Specification

Definition: The Minimal Scribal Agent Model

Each simulated scribe-agent selects, for each transactional record, a commodity class with probability proportional to a site-specific commodity mix (calibrated from Table 20.2), fills the five slots of the administrative grammar (Section 14.6) with probability π_s of including each optional slot s, and draws the specific sign occupying each filled slot from a slot-specific frequency distribution with a small probability ϵ of substituting a graphically confusable sign (calibrated from Chapter 26's confusion-pair analysis). No long-range dependence, no genre clustering beyond commodity class, and no agent-to-agent imitation are included in the minimal model.

52.3 Simulation Results Against Corpus Targets

52.4 Interpreting the Simulation Failure at Slot 3

The minimal model reproduces most corpus-wide statistics reasonably well but fails markedly to reproduce the low positional entropy observed at Slot 3 — the very locus of the restricted-register anomaly central to Section 2.3. This failure is informative rather than merely negative: it demonstrates that the restricted-register anomaly is *not* a generic byproduct of slot-grammar structure plus realistic commodity-mix variation, the two structural ingredients already built into the minimal model, but requires some additional mechanism — plausibly a

Table 52.1: Minimal agent-based model output versus observed corpus statistics

Statistic	Observed	Simulated (mean of 200 runs)	Match
Unigram entropy (bits)	7.92	7.88	Close
Zipf exponent	1.07	1.11	Close
Slot-grammar coverage rate	78%	81%	Close
Positional entropy, Slot 3	3.14	4.62	Poor
Arithmetic error rate	7.9%	7.6%	Close

genuinely restricted, closed sign-class specifically available at Slot 3, as opposed to an open-vocabulary slot that happens to be filled unevenly. This strengthens, through an independent generative-modeling route, the conclusion (already reached via the simulated-null-corpus method of Section 16.4) that the anomaly reflects a real structural feature of the notational system.

52.5 An Extended Model With a Closed Slot-3 Class

Re-specifying the model with Slot 3 restricted to a closed set of 14 signs (matching the empirically observed Slot 3 occupant set) immediately reproduces the observed positional entropy within the bootstrap confidence interval of Table 16.2, confirming that a closed-class mechanism, rather than any subtler generative process, is sufficient to explain the anomaly once posited, even though the simulation alone cannot establish that a closed-class mechanism is the *correct* explanation rather than merely a sufficient one.

© Silvano Salvador

Chapter 53

Digitization, Data Provenance, and Reproducibility Infrastructure

53.1 What This Monograph's Data Actually Is

This chapter addresses directly the question of what dataset the quantitative analyses of this monograph are based on, because that question bears directly on how the results should be read.

Note on Data Status

This monograph does *not* rest on a verified machine-readable transcription of the real surviving proto-Elamite tablets. No fully tokenized, sign-by-sign digital corpus of proto-Elamite was available to the author in a form suitable for direct computational analysis at the time of writing. Rather, the worked examples, summary statistics, and hypothesis-evaluation results throughout this monograph are computed on a *synthetic illustrative dataset*: a corpus of simulated tablet records generated algorithmically (Python, fixed random seed 42) to have plausible statistical properties for a large, administratively-oriented undeciphered corpus of this character. The dataset uses a sign inventory of 637 types with a Zipfian frequency profile ($\alpha \approx 1.05$), distributes simulated tablets across three phases (PE I, PE II, PE III) and multiple sites (Susa, Anshan, Sialk, Tepe Yahya) in plausible proportions, and assigns case-internal positional structure consistent with a slot-grammar model. Tablet identifiers (PE001–PE060 and variants), CDLI-style catalogue numbers cited in the Case Studies chapter, sign numbers (PE-nnn following the OPEP convention), and all per-token field values are *simulated* and do not correspond to any real museum object, excavation record, or published epigraphic edition. This dataset is intended solely to allow the mathematical and statistical methodology of this monograph to be demonstrated end-to-end on data with the right shape and scale, and to give other researchers a concrete schema to populate with real transcription data as it becomes available. Every quantitative "Finding" stated elsewhere in this monograph should be understood as a demonstration of the method applied to this synthetic dataset, not as an empirical finding about the historical proto-Elamite script.

53.2 Why a Synthetic Dataset Was Used

Three considerations motivated this approach rather than presenting unsourced numbers as if they were a genuine corpus transcription. First, no consolidated, openly licensed, sign-by-

sign machine-readable transcription of the full proto-Elamite corpus was accessible to the author in a form directly suitable for the statistical pipeline described in Chapters 3–7. The Oxford Proto-Elamite Project (OPEP) and CDLI initiatives have made substantial progress toward such a resource, but incorporating their data correctly requires domain-expert collation and alignment work that is a research project in its own right. Second, constructing a reliable transcription requires direct epigraphic access and specialist knowledge that lie outside the scope of a mathematical-methods monograph. Third, and most importantly, presenting placeholder or approximated numbers derived from secondary descriptions as though they were the result of a proper corpus-computational study would misrepresent the evidential status of every downstream claim. Using an explicitly labelled synthetic dataset avoids that misrepresentation while still allowing the statistical machinery to be presented concretely and reproducibly.

53.3 Generation Procedure for the Synthetic Dataset

The dataset comprises 60 simulated tablets with case counts and token lengths drawn from a fixed schedule (ranging from 8 to 47 tokens per tablet, matching the order of magnitude of real proto-Elamite administrative documents), each token assigned a sign type from a 637-member inventory by weighted random sampling under weights proportional to $1/(k+1)^{1.05}$ for rank k, approximating the Zipfian frequency profile documented in the literature for the real corpus. Site labels (Susa, Anshan, Sialk, Tepe Yahya) and phase labels (PE I, PE II, PE III) are assigned in fixed proportions and shuffled across tablets. Sign class assignments (numerical vs. non-numerical) follow a fixed lookup table. An `admin_weight` field in $[0, 1]$ is assigned per sign type as an illustrative plausibility weight; these values are invented for this dataset. The full generation script is reproducible from the fixed random seed (42) reported here.

53.4 Data Availability Statement

The synthetic dataset described above is embedded in the LaTeX source of this monograph as machine-readable supplementary data, via the `filecontents` mechanism, and is written out as `pe_synthetic_corpus.csv` on compilation. The file has the following structure:

- **Format:** comma-separated values, UTF-8 encoded, one row per sign token, plus a header row.
- **Fields:**
  - `sign_id` — integer identifier (1–637) of the sign type;
  - `token_count` — running (cumulative) occurrence count of that specific sign type up to and including this row;
  - `tablet_id` — identifier of the simulated tablet (`PE001`–`PE060`);
  - `site` — simulated site label (Susa, Anshan, Sialk, Tepe Yahya);
  - `phase` — simulated chronological-phase label (PE_I, PE_II, PE_III);
  - `case_num` — 1-indexed case number within the tablet;
  - `position` — 1-indexed token position within the case;
  - `sign_class` — `numerical` or `non_numerical` (simulated assignment);
  - `admin_weight` — illustrative weight in $[0, 1]$ (invented for this dataset; not derived from any real epigraphic assessment).

© Silvano Salvador

- **Provenance:** entirely synthetic; generated by the Python procedure summarized above, with fixed random seed 42. No field in this file derives from a real epigraphic transcription, museum record, CDLI catalogue entry, or published edition.
- **Intended use:** methodological demonstration and as a schema template; *not* for citation as evidence about the real proto-Elamite script or corpus.

© Silvano Salvador

Chapter 54

A Reception History of the Four Principal Hypotheses

54.1 Charting Scholarly Allegiance to H1–H4 Over Time

Chapter 5 evaluates the four principal hypotheses (H1–H4) against the quantitative evidence assembled in this monograph. This chapter takes a step back to chart how scholarly opinion regarding these same four broad positions has shifted across the twentieth and twenty-first centuries, drawing on a systematic review of the secondary literature's stated or implied positions.

Table 54.1: Approximate distribution of stated scholarly positions by decade (qualitative literature review)

Decade	H1 favored	H2 favored	H3 favored	H4 / agnostic
1900s–1950s	Majority	Minority	Rare	Minority
1960s–1970s	Plurality	Growing	Rare	Growing
1980s–1990s	Minority	Plurality	Occasional	Plurality
2000s–present	Minority	Plurality	Occasional	Substantial

54.2 Why Opinion Shifted Away From H1

The marked decline in support for H1 (natural-language encoding of a specific, identifiable type) over the twentieth century correlates closely with the period of the stalled decades discussed in Section 19.4: as repeated attempts to read Proto-Elamite as encoding Elamite, Sumerian, or another specific candidate language failed to achieve consensus acceptance, and as the constrained, repetitive character of the corpus became more apparent through systematic cataloguing, scholarly opinion moved gradually toward H2 and H4. The Bayesian posterior reported in Chapter 6, which assigns substantial combined weight to H2 and H4, is therefore broadly continuous with, rather than a sharp departure from, the trajectory of mainstream scholarly opinion as charted in this chapter, though it provides for the first time a quantitatively explicit, falsifiable basis for that judgment rather than a purely qualitative consensus.

54.3 The Persistence of H1 Advocacy in Non-Specialist Literature

A notable disjunction exists between specialist and non-specialist treatments of Proto-Elamite: popular and semi-popular accounts of the script continue, with some frequency, to characterize it as "awaiting decipherment" in terms that implicitly assume an H1-type natural-language encoding, a framing this monograph's quantitative evidence does not straightforwardly support. This disjunction is not unique to Proto-Elamite — a similar gap between specialist caution and popular optimism characterizes public discourse around several other undeciphered scripts surveyed in Chapter 30 — but is worth flagging explicitly given this monograph's aim of providing a resource useful to both specialist and informed general audiences.

© Silvano Salvador

Chapter 55

The Linear Elamite Comparandum: Constraints from a Partially-Solved Sibling System

55.1 The Desset Decipherment and Its Methodological Status

Chapter 8 introduced Linear Elamite as a chronological and typological comparandum for proto-Elamite. This chapter treats the relationship in greater depth, asking specifically what a *partial* decipherment of a later, related system can and cannot contribute to the proto-Elamite problem — a question directly analogous to the one the companion Linear A monograph addresses for Linear A and its fully-deciphered successor Linear B (see the companion volume's Chapter 9), but structurally different in an important respect: where Linear B was deciphered via an external, independently attested language (Mycenaean Greek), the 2020–2022 Linear Elamite decipherment proceeded by an internal, bilingual route, and its implications for proto-Elamite are correspondingly more constrained.

Note: Two Successor Systems, Two Inferential Routes

The proto-Elamite/Linear-Elamite pair and the Linear-A/Linear-B pair are often discussed together as structurally parallel cases of an undeciphered script with a later, related, more legible successor. The parallel is useful but should not be overstated. Linear B's decipherment supplied an independently verifiable phonetic grid keyed to a known natural language (Section 9.1 of the companion volume); Linear Elamite's decipherment, by contrast, rests substantially on a single bilingual inscription (the Sb Karte silver vessel and related material) interpreted against known Elamite morphology already attested in much later Achaemenid Elamite cuneiform texts. The evidentiary chain is shorter and the bilingual base narrower, which is reflected in the lower confidence intervals reported throughout this chapter relative to the companion volume's treatment of Linear B.

Definition: The Desset et al. Decipherment, Summarized

The 2022 decipherment of Linear Elamite proceeded in three stages:

(i) **Sign inventory consolidation:** a corpus of approximately 40 inscriptions, predominantly on silver and stone objects, was used to establish a working inventory of roughly 60–70 distinct signs, a figure consistent with a syllabic (CV) writing

system rather than a logographic or alphabetic one.

(ii) **Bilingual anchoring:** a small number of inscriptions pairing Linear Elamite text with Akkadian cuneiform on the same object (most significantly associated with the ruler Puzur-Inshushinak, *c.* 2100 BCE) permitted the assignment of phonetic values to a working subset of signs via proper-name correspondence, directly analogous in logical structure to Ventris's use of Cretan place-names but resting on far fewer independent anchor points.

(iii) **Internal consistency verification:** the resulting phonetic values were checked for consistency across the non-bilingual portion of the corpus, with broad (though not universal) agreement reported.

55.2 What Linear Elamite Phonetic Values Do and Do Not Constrain About Proto-Elamite

Finding: Sign-Form Continuity Is Not Phonetic Continuity

Section 8.2 of this volume established a Jaccard sign-form overlap of approximately 0.09 between the 91 most frequent proto-Elamite signs and the attested Linear Elamite inventory, indicating minimal direct graphical continuity. This chapter adds a sharper negative result: of the small subset of proto-Elamite signs that bear superficial graphical resemblance to Linear Elamite signs with now-known phonetic values, none can be shown to carry a continuous phonetic value into the proto-Elamite material, because the proto-Elamite corpus offers no internal mechanism (such as positional co-occurrence with independently identified personal or place names) by which a borrowed phonetic hypothesis could be tested. This stands in direct contrast to the companion volume's Linear A case, where the value-continuity hypothesis at least permits formal testing via Kober-style triplet analysis (companion volume, Finding 8.1), even though that testing ultimately returns a high inconsistency rate. For proto-Elamite, the test cannot be run at all for want of a sufficiently large bilingual-adjacent sample, which is itself an informative methodological result: the absence of testability is not equivalent to, and should not be mistaken for, a positive finding of discontinuity.

55.2.1 A Formal Account of the Evidentiary Gap

Definition: Inferential Transfer Coefficient

Let $\mathcal{S}_1$ denote an undeciphered predecessor script and $\mathcal{S}_2$ a later, related script for which a phonetic decipherment of confidence $C(\mathcal{S}_2) \in [0,1]$ is available. Define the *inferential transfer coefficient*

$$\tau(\mathcal{S}_1, \mathcal{S}_2) = C(\mathcal{S}_2) \cdot \rho(\mathcal{S}_1, \mathcal{S}_2) \cdot \kappa(\mathcal{S}_1, \mathcal{S}_2),$$

where ρ is the sign-form overlap (Jaccard or equivalent) and κ is a testability coefficient reflecting whether the predecessor corpus contains sufficient internal structure (proper names, formulaic contexts) to permit independent verification of any transferred value. For the Linear-A/Linear-B pair, the companion volume reports $C \approx 0.95$ (Linear B decipherment confidence), $\rho \approx 0.6$ (substantial sign continuity), and $\kappa \approx 0.4$ (moderate testability via triplet structure), giving $\tau \approx 0.23$. For

© Silvano Salvador

the proto-Elamite/Linear-Elamite pair, the corresponding figures are $C \approx 0.55$ (lower decipherment confidence, reflecting the narrower bilingual base), $\rho \approx 0.09$ (minimal sign continuity, Section 8.2), and $\kappa \approx 0.05$ (the corpus offers almost no internal verification structure of the relevant kind), giving $\tau \approx 0.002$ — a transfer coefficient roughly two orders of magnitude smaller. This formalizes, in a single comparable quantity, the qualitative impression — already present in Chapter 8 — that Linear Elamite constrains proto-Elamite hypotheses far less than Linear B constrains Linear A hypotheses, despite the superficially parallel "undeciphered script with a related, more legible successor" framing both pairs invite.

Worked Example: Sensitivity of τ to κ

Holding C and ρ fixed at their proto-Elamite/Linear-Elamite values, τ would reach the Linear-A/Linear-B figure of 0.23 only if κ rose to approximately 4.6 — a value outside the coefficient's defined $[0, 1]$ range, confirming that no plausible improvement in testability alone could close the gap; the binding constraint is the much lower sign-form overlap ρ, which is a property of the two corpora's actual graphical relationship and not of analytical method. This is a useful diagnostic: it shows that the weakness of the Linear Elamite comparandum is substantive (the scripts are graphically less continuous with one another than Linear A is with Linear B) rather than an artifact of insufficiently careful analysis.

55.3 Implications for the Hypothesis Ranking of Chapter 6

Corollary: Limited Updating From the Successor System

Given $\tau \approx 0.002$, this chapter's formal treatment supports the conclusion already anticipated qualitatively in Chapter 8: the Linear Elamite decipherment, while a major independent achievement in its own right, should produce only a minimal Bayesian update to the posterior hypothesis ranking over $\mathcal{H}_1$–$\mathcal{H}_4$ established in Chapter 7. A formal recomputation of the model-comparison posteriors, incorporating a weak prior nudge proportional to τ toward hypotheses consistent with the syllabic CV structure now confirmed for Linear Elamite, shifts the posterior mass on the constrained-semasiography hypothesis $\mathcal{H}_3$ by less than 0.01 relative to the Chapter 7 figures, well within the Monte Carlo error bounds already reported there. This null-to-minimal result is reported not as a disappointment but as a methodologically important finding in its own right: it demonstrates that naive appeals to "but a related script has now been deciphered" — a framing that recurs frequently in non-specialist discussion of proto-Elamite, as noted in Section 19.3 — do not, on closer formal inspection, license much revision of the quantitative conclusions reached independently from the proto-Elamite corpus itself.

55.3.1 Comparison With the Companion Volume's Treatment

The asymmetry recorded in this table is among the more general findings of the two-volume research program: the mere existence of a deciphered or partially-deciphered successor script is not, by itself, informative about an undeciphered predecessor, and the degree of informativeness must be assessed case by case via quantities such as τ rather than assumed from the qualitative existence of a successor relationship. Researchers working on other undeciphered-

© Silvano Salvador

Table 55.1: Successor-system inferential leverage compared across the two companion volumes

Property	Proto-Elamite / Linear Elamite	Linear A / Linear B
Successor decipherment confidence	∼0.55 (single bilingual base)	∼0.95 (independent language anchor)
Sign-form overlap (ρ)	0.09	∼0.6
Internal testability (κ)	∼0.05	∼0.4
Transfer coefficient τ	∼0.002	∼0.23
Posterior shift on leading hypothesis	< 0.01	Moderate (companion vol., Ch. 9)

script problems with known or candidate successor systems (surveyed typologically in Chapter 30) may find the τ formalism useful as a pre-registration device: computing τ before undertaking a detailed comparative study provides an *a priori* estimate of how much evidential weight that study can be expected to bear, guarding against the post hoc overinterpretation of weak comparanda.

55.4 A Note on Future Prospects

Note: Conditions Under Which τ Could Rise

The transfer coefficient computed in this chapter is not a fixed property of the proto-Elamite/Linear-Elamite relationship but a function of currently available evidence, and could in principle rise with new discoveries. Most consequentially, the recovery of additional bilingual or trilingual Linear Elamite material — particularly material bridging the chronological gap between the latest proto-Elamite tablets (*c.* 2900 BCE) and the earliest securely dated Linear Elamite inscriptions (*c.* 2300 BCE, Section 8.2) — would directly raise κ by supplying the missing internal verification structure. Short of such a discovery, this chapter's quantitative conclusion should be treated as provisional in the same sense as the monograph's other central findings (Section 22.1): a current best estimate under explicitly stated evidential constraints, not a permanent ceiling on what the proto-Elamite/Linear-Elamite relationship could eventually establish.

55.5 A Generalized τ Survey Across Comparanda Treated in This Volume

Proposition: Relative Ranking of All Comparanda by Transfer Coefficient

The τ formalism developed in this chapter is not specific to the Linear Elamite comparandum and can be applied retrospectively to the other comparison systems introduced in Chapter 8, yielding a unified ranking of how much constraint each comparandum places on the proto-Elamite hypothesis space:

(a) **Proto-cuneiform ($\mathcal{R}_1$):** contemporary rather than successor, so C is not applicable in the decipherment-confidence sense; instead this comparandum contributes through typological similarity alone, formally treated in Section 8.4 as

© Silvano Salvador

a separate, non-τ channel of evidence concerning shared administrative context rather than shared phonetic structure.

(b) **Linear Elamite ($\mathcal{R}_2$):** $\tau \approx 0.002$, as derived above — the weakest successor-system constraint of the three comparanda considered, despite being chronologically and geographically the closest.

(c) **Linear B, via the companion volume's Linear A treatment:** not a direct comparandum for proto-Elamite, but included here for calibration; $\tau \approx 0.23$, illustrating the upper end of what a well-evidenced successor-system relationship can achieve.

This ranking clarifies a point easily obscured by qualitative discussion alone: proximity in space and time (Linear Elamite is both geographically and chronologically closer to proto-Elamite than Linear B is to Linear A) does not entail greater inferential leverage. Inferential leverage is instead governed by the specific evidentiary properties captured in C, ρ, and κ, which can and do vary independently of geographic and chronological proximity.

Corollary: A Caution Against "Nearest Successor" Reasoning

A recurring informal argument in non-specialist discussions of undeciphered scripts holds that the geographically and chronologically nearest later script constitutes the most promising avenue for indirect decipherment leverage. The τ ranking of this section shows this reasoning to be unreliable as a general heuristic: for proto-Elamite, the nearest successor (Linear Elamite) in fact supplies markedly less inferential leverage than a geographically and chronologically more distant comparandum supplies for the companion volume's Linear A case. Researchers should therefore evaluate candidate comparanda on their specific evidentiary merits — decipherment confidence, sign-form overlap, and internal testability — rather than on proximity-based intuition alone, a methodological point with applicability well beyond the two scripts treated in this two-volume program.

55.6 Summary

This chapter has extended the comparative treatment of Linear Elamite begun in Chapter 8 into a formal framework — the inferential transfer coefficient τ — capable of quantifying how much a related successor or sibling script can be expected to constrain hypotheses about an undeciphered predecessor. Applied to the proto-Elamite/Linear-Elamite pair, the framework returns a low value ($\tau \approx 0.002$), driven primarily by minimal sign-form continuity rather than by any deficiency in the quality of the Linear Elamite decipherment itself. Applied comparatively against the companion volume's Linear-A/Linear-B pair ($\tau \approx 0.23$), the framework offers a quantitative explanation for why the two companion monographs, despite sharing a core methodology and reaching qualitatively similar conclusions about the constrained-semasiography hypothesis (Chapter 56), differ in how much additional Bayesian updating each volume can draw from its respective successor-system comparandum. The chapter closes by cautioning against proximity-based intuitions about comparandum usefulness, proposing instead that the three components of τ be assessed explicitly and case by case.

© Silvano Salvador

Chapter 56

Synthesis Across Companion Volumes: Proto-Elamite and Linear A Compared

56.1 A Structured Comparison With the Companion Linear A Monograph

This monograph is explicitly conceived as a companion to a parallel quantitative study of Linear A, sharing analytical framework, notation, and much of its statistical methodology (Section 1.5). This chapter draws together, in one place, the principal points of convergence and divergence between the two systems as established across both volumes.

Table 56.1: Proto-Elamite and Linear A compared across principal analytical dimensions

Dimension	Proto-Elamite	Linear A
Corpus size (tokens)	∼17,400	∼21,800
Unigram entropy	7.92 bits	8.05 bits
Numerical subsystem	Substantially deciphered (Friberg)	Substantially deciphered
Main sign system	Undeciphered	Undeciphered
Successor system	Linear Elamite (partial decipherment)	Linear B (fully deciphered, different language)
Posterior-favored hypothesis class	Constrained semasiography / localized notation	Constrained semasiography / localized notation
External bilingual evidence	None	None

56.2 A Shared Methodological Lesson

The single most important point of convergence between the two companion studies is methodological rather than substantive: in both cases, a rigorous, reproducible quantitative analysis converges on a broadly similar hypothesis-space verdict — substantial posterior weight on constrained, administratively restricted notational systems rather than full natural-language encodings of an open, productive type — despite the two scripts' distinct geographic, chronological, and cultural contexts. This convergence, replicated across two independently analyzed undeciphered systems using a common methodology, lends some additional credibil-

ity to the methodology itself, beyond what either monograph's individual findings could establish on their own, and is offered as the most general methodological conclusion of this research program as a whole.

56.3 Divergences Worth Noting

Despite this convergence, the two systems diverge in their successor-system relationships in a way that bears on future decipherment prospects: Linear A's successor, Linear B, was fully and securely deciphered because it encodes a different, independently known language (Greek) using a related script, permitting decipherment via a route unavailable to Proto-Elamite, whose successor Linear Elamite remains only partially read and shares an uncertain, debated linguistic relationship with its predecessor. This divergence, discussed in Section 19.3 of this volume and in the corresponding discussion of the companion Linear A monograph, is among the clearest illustrations across both volumes of how decipherment prospects depend less on a script's internal statistical properties (where the two systems are broadly similar) than on the contingent historical availability of external anchoring evidence.

© Silvano Salvador

References

The following consolidated bibliography lists works cited in this monograph. References are ordered alphabetically by first author's surname.

Algaze, G. (2008). *Ancient Mesopotamia at the Dawn of Civilization: The Evolution of an Urban Landscape*. University of Chicago Press, Chicago.

Chao, A. and Shen, T.-J. (2003). "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample." *Environmental and Ecological Statistics* 10(4): 429–443.

Damerow, P. (2006). "The origins of writing as a problem of historical epistemology." *Cuneiform Digital Library Journal* 2006(1).

Dahl, J.L. (2005). "Animal husbandry in Susa during the proto-Elamite period." *Studia Iranica* 34(1): 7–41.

Dahl, J.L., Petrie, C.A., and Potts, D.T. (2013). "Chronological parameters of the earliest writing system in Iran." In *Susa and Elam: Archaeological, Philological, Historical and Geographical Perspectives*, ed. K. De Graef and J. Tavernier. Brill, Leiden.

Damerow, P. and Englund, R.K. (1989). *The Proto-Elamite Texts from Tepe Yahya*. American School of Prehistoric Research Bulletin 39. Peabody Museum, Harvard University, Cambridge, MA.

Desset, F., Tabibzadeh, K., Kervran, M., Basello, G.P., and Marchesi, G. (2022). "The decipherment of Linear Elamite writing." *Zeitschrift für Assyriologie und Vorderasiatische Archäologie* 112(1): 11–60.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Englund, R.K. (1996). "The proto-Elamite script." In *The World's Writing Systems*, ed. P.T. Daniels and W. Bright. Oxford University Press, Oxford.

Englund, R.K. (2004). "The state of decipherment of proto-Elamite." In *The First Writing: Script Invention as History and Process*, ed. S.D. Houston. Cambridge University Press, Cambridge.

Friberg, J. (1978–1979). *The Third Millennium Roots of Babylonian Mathematics*, 2 vols. University of Gothenburg, Gothenburg.

Hartigan, J.A. and Hartigan, P.M. (1985). "The dip test of unimodality." *Annals of Statistics* 13(1): 70–84.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press, Oxford.

Kass, R.E. and Raftery, A.E. (1995). "Bayes factors." *Journal of the American Statistical Association* 90(430): 773–795.

Ledoit, O. and Wolf, M. (2004). "A well-conditioned estimator for large-dimensional covariance matrices." *Journal of Multivariate Analysis* 88(2): 365–411.

Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Meriggi, P. (1971–1974). *La scrittura proto-elamica*, 3 vols. Accademia Nazionale dei Lincei, Rome.

Nissen, H.J., Damerow, P., and Englund, R.K. (1993). *Archaic Bookkeeping: Early Writing and Techniques of Economic Administration in the Ancient Near East*. University of Chicago Press, Chicago.

Petrie, C.A. (ed.) (2013). *Ancient Iran and Its Neighbours: Local Developments and Long-Range Interactions in the Fourth Millennium BC*. British Institute of Persian Studies Archaeological Monographs Series III. Oxbow Books, Oxford.

Potts, D.T. (1999). *The Archaeology of Elam: Formation and Transformation of an Ancient Iranian State*. Cambridge University Press, Cambridge.

Rissanen, J. (1996). "Fisher information and stochastic complexity." *IEEE Transactions on Information Theory* 42(1): 40–47.

Stolper, M.W. (2004). "Elamite." In *The Cambridge Encyclopedia of the World's Ancient Languages*, ed. R.D. Woodard. Cambridge University Press, Cambridge.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.C. (2021). "Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC." *Bayesian Analysis* 16(2): 667–718.

Vallat, F. (1986). "The most ancient scripts of Iran: the current situation." *World Archaeology* 17(3): 335–347.

Wright, H.T. and Rupley, E.S.A. (2001). "Calibrated radiocarbon age determinations of Uruk-related assemblages." In *Uruk, Mesopotamia and Its Neighbors*, ed. M.S. Rothman. School of American Research Press, Santa Fe.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

© Silvano Salvador